Florian Heinke

Energieprofilbasierende Analysemethoden von Proteinfamilien

BACHELORARBEIT

HOCHSCHULE MITTWEIDA

UNIVERSITY OF APPLIED SIENCES

Mathematik, Naturwissenschaften, Informatik

Mittweida, 2010

Florian Heinke

Energieprofilbasierende Analysemethoden von Proteinfamilien

eingereicht als

BACHELORARBEIT

an der

HOCHSCHULE MITTWEIDA

UNIVERSITY OF APPLIED SIENCES

Mathematik/Naturwissenschaften/Informatik

Mittweida, 2010

Erstprüfer: Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer: Dipl.-Inf. (FH) Daniel Stockmann

Bibliografische Beschreibung

Heinke, Florian:
Energieprofilbasierende Analysemethoden von Proteinfamilien. – 2010. – 83 S.
Mittweida, Hochschule Mittweida,
Fakultät Mathematik/Naturwissenschaften/Informatik,
Bachelorarbeit, 2010

Referat

Als eines der wichtigsten Verfahren der Biologie als beobachtende Wissenschaft ist das Vergleichen von Organismen. Zweck dafür ist es auf phylogenetische Zusammenhänge schließen zu können. Dieses Prinzip ist ebenso auf den Nanokosmos der Proteine übertragbar. Aus dem Vergleich zweier Sequenzen können Unterschiede und Ähnlichkeiten aufgedeckt, und somit Rückschlüsse auf die funktionellen, strukturellen und evolutionären Beziehungen gewonnen werden.

Es existieren zahlreiche Algorithmen, die den Vergleich von Proteinen auf den verschiedensten Abstraktionsebenen ermöglichen, wobei diese meist hochgradig spezialisiert sind. Diese Determiniertheit führt häufig zum Informationsverlust. Um den biologischen Kontext zu erfassen, ist es oft notwendig verschiedene Algorithmen auf eine Fragestellung anzuwenden.

In dieser Arbeit wird ein Algorithmus vorgestellt, der die Informationen verschiedenster Proteinstrukturebenen vereint und daraus ein einziges Alignment erzeugen kann. Dafür nutzt das Programm den neuartigen Ansatz der Proteinenergieprofilberechnung.

Weiterhin wird im Rahmen dieser Arbeit näher auf die Berechnung und Strukturkorrelationen von Energieprofilen eingegangen. Zudem wird ein Algorithmus erläutert, der Alignments dieser Profile durchführt. Im letzten Punkt wird eine Methodik zur Vorhersage von Energieprofilen erläutert sowie dessen Güte abgeschätzt.

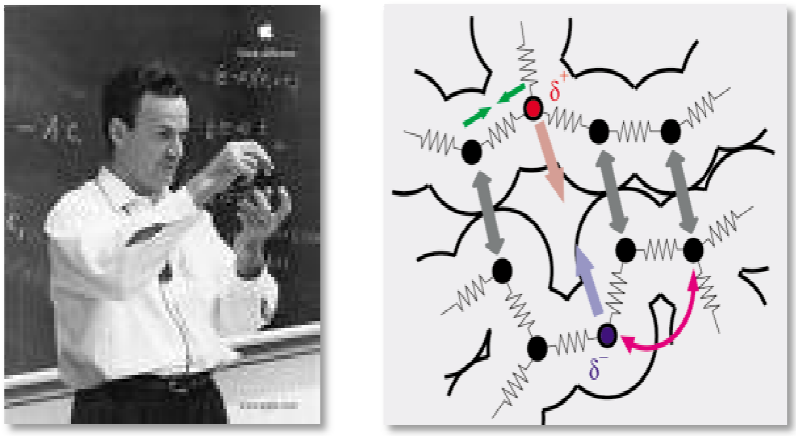
Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung/Motivation

Richard Feynman prognostizierte im Jahr 1959 als erster Wissenschaftler die Zukunft der Nanotechnologie und äußerte die Vermutung, dass Maschinen und Materialien eines Tages auf atomarer Ebene konstruiert werden könnten: „Die physikalischen Gesetze sprechen, soweit ich das beurteilen kann, nicht dagegen, Dinge Atom für Atom bewegen zu können.“



…en (rechts) [1]

…egt werden könnten

Dieser Satz inspirierte mich, über einen Zusammenhang zwischen Biologie und Mechanik nachzudenken. Ist es möglich Proteine und andere Biomoleküle auf der Ebene von Kräften und der damit verbundenen Energie darzustellen und den Zusammenhang zu ihrer Struktur rein mechanisch abzubilden? Ist es möglich Proteine auf Grund von Energie und Kräften, die zwischen ihren Atomen vorhanden sind, zu vergleichen und sogar vorherzusagen?

Proteine sind Makromoleküle, die aufgrund ihrer katalytischen Eigenschaften in der Lage sind, lebenswichtige Stoffwechselvorgänge zu realisieren. Im Laufe der Evolution haben sich diese „Nanofabriken“ differenziert und dabei eine enorme Struktur- und Funktionsvielfalt ausgebildet.

Das Verstehen dieser Strukturen und Funktionalitäten der Proteine gehört zu den bedeutendsten Zielen der modernen Biologie bzw. zu den größten Herausforderungen der *life siences*. Die Erkenntnisse der letzten 15 Jahre hatten bereits eklatante

Auswirkungen auf die Pharmakologie und auf das stetig wachsende Feld der BioNanotechnologie. Zum einen können Medikamente in ihrer Wirkungsweise optimiert als auch weitaus schneller entwickelt und hergestellt werden. So ermöglichen computergestützte Technologien der Bioinformatik unter Verwendung von Protein- und DNA-Datenbanken die Identifizierung von krankheitsinduzierenden Proteinen sowie für den Verlauf der Krankheit notwendige Proteine (*targets*) in relativ kurzer Zeit.

Ein solches *target* ist die Protease des HI-Virus. Dieses Protein ist essenziell für den Replikationszyklus des Virus. Neuste Studien zeigten, dass eine Deaktivierung (Inhibierung) dieses Proteins mittels verschiedenster chemischer Substanzen, exemplarisch TL-3, möglich ist, und somit die Verbreitung des Virus gestoppt und das Ausbrechen der Krankheit bei bereits infizierten Menschen unterbunden werden könnte [2][3].

Eine weitere Möglichkeit der BioNanotechnologie wird es sein, künstliche Proteine zu synthetisieren, die in ihrer Funktion, Struktur und chemischen Stabilität an die gewünschte Aufgabe angepasst sind.

Letzteres erweist sich bis jetzt als unmöglich, da das Wissen über Struktur und Funktionalität nicht ausreichend ist. Ein wesentliches Ziel der Bioinformatik ist es, die Lücke zwischen der Sequenz eines Proteins und dessen Struktur sowie Funktion weiter zu schließen.

In dieser Arbeit werden neuartige Algorithmen der Proteinanalytik vorgestellt. Basis hierfür bildet der noch junge Ansatz der Berechnung von Protein-Energieprofilen [4]. Zum einen wird die Entwicklung und Durchführung eines so genannten Super-Alignments dargestellt. Dieses führt die Informationen der verschiedenen Abstraktionsebenen zweier Proteine zusammen. Aus dieser Vereinigung von Sequenz, Sekundärstruktur und Energie, entsteht ein mehrdimensionales Alignment.

Weiterhin wird auf den Zusammenhang zwischen Energie und Proteinstruktur eingegangen. Abschließend werden Möglichkeiten zum Vergleich und Vorhersage von Energieprofilen dargelegt.

2. Theoretische Grundlagen

2.1 Proteine

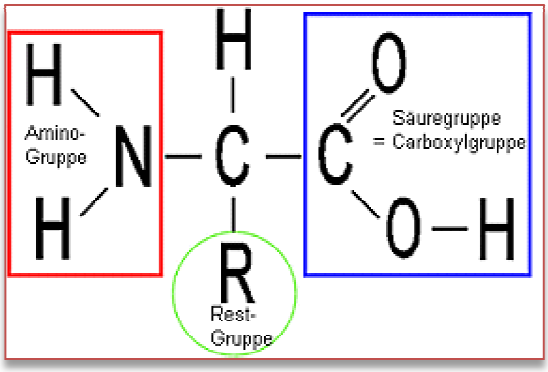
Proteine sind lineare Makromoleküle, die sich aus den 20 natürlich vorkommenden Aminosäuren zusammensetzen. Sie sind ein wesentlicher Bestandteil jeder Zelle, da sie notwendige Stoffwechselvorgänge realisieren. Im Laufe der Evolution hat sich in Prokaryoten sowie Eukaryoten eine enorme Fülle von Proteinen entwickelt. Entsprechend hat sich die Funktionalität differenziert [7].

In erster Linie kann man zwischen globulären, also den frei im Cytoplasma bzw. dem entsprechenden Zellorganell befindlichen Proteinen, und den Membranproteinen unterscheiden. Letztere sind für Pharmazie bzw. Virologie besonders interessant, da diese Eiweiße potentielle Targets für den aktiven Transport von antibiotische bzw. biostatische Substanzen durch die Zellmembran sind. Transmembranproteine, eine spezielle Gruppe der Membranproteine, wirken stoffselektiv, d.h. Moleküle mit einer bestimmten Größe sowie die meisten Ionen werden nicht in die Zelle transportiert [5]. Leider ist nur wenig über diese Proteingruppe bekannt. Zwar ließen sich mittels Edman-Abbau die Primärstrukturen etlicher Membranproteine sequenzieren, die räumliche Struktur hingegen konnte mittels der klassischen Methoden der NMR und Kristallröntgenspektroskopie nur in wenigen Fällen ermittelt werden. Grund für diese Schwierigkeit ist, dass Membranproteine fest in der Membran integriert sind [7, 15]. Löst man sie für analytische Zwecke aus diesem Milieu, verändert sich die räumliche Struktur und das Ergebnis ist verfälscht bzw. wertlos. Anders verhält es sich bei globulären Proteinen. Wie bereits beschrieben, befinden sich diese Eiweiße frei in einem wässrigen Milieu. Dadurch ist es leichter die Proteine aus dem entsprechenden Zellorganell zu extrahieren und zu kristallisieren [15]. Zum gegenwärtigen Zeitpunkt sind in der Protein Data Bank (PDB) ca. 67300 Strukturen aufgeführt. Bei ungefähr 4900 Eiweißen handelt es sich um Membran-Proteine, was einen Anteil von nur 7 % ausmacht [16, 17]. Aufgrund der geringen Anzahl von Membranproteinstrukturen liegt das Hauptaugenmerk dieser Arbeit auf globulären Proteinen.

2.2 Aufbau und Struktur der Proteine

2.2.1 Primär- und Sekundärstruktur

Wie bereits im vorangegangen Punkt beschrieben, handelt es sich bei Proteinen um lineare Makromoleküle die sich, bis auf wenige Ausnahmen, aus den 20 kanonischen Aminosäuren zusammensetzen.



, wobei sie sich

Wie die Abbildung 2 zeigt, lassen sich alle biologisch relevanten Aminosäuren auf einen grundlegenden Aufbau differenzieren. Zum einen besitzt jede Aminosäure mindestens eine Aminogruppe und einer Carboxylgruppe. Letztere funktionelle Gruppe wirkt H^+-donierend, d.h. das an den Sauerstoff gebundene Wasserstoffatom wird an die Umgebung abgegeben, wo es als Proton vorliegt und entsprechend reagieren kann. Die Aminogruppe hingegen ist in der Lage in der Umgebung vorkommende Protonen aufzunehmen.

Die chemischen Eigenschaften werden allerdings durch die Restgruppe determiniert. Diese ist am ersten Kohlenstoffatom der Hauptkette, dem so genannten α-C-Atom, kovalent gebunden. Die Struktur dieser Restgruppe bestimmt relevante Charakteristika, wie Polarität, Hydrophobizität sowie das Säure- oder Base-Verhalten [5], sodass eine Klassifikation mittels dieser chemisch-physikalischen Eigenschaften möglich ist (Abbildung 3).

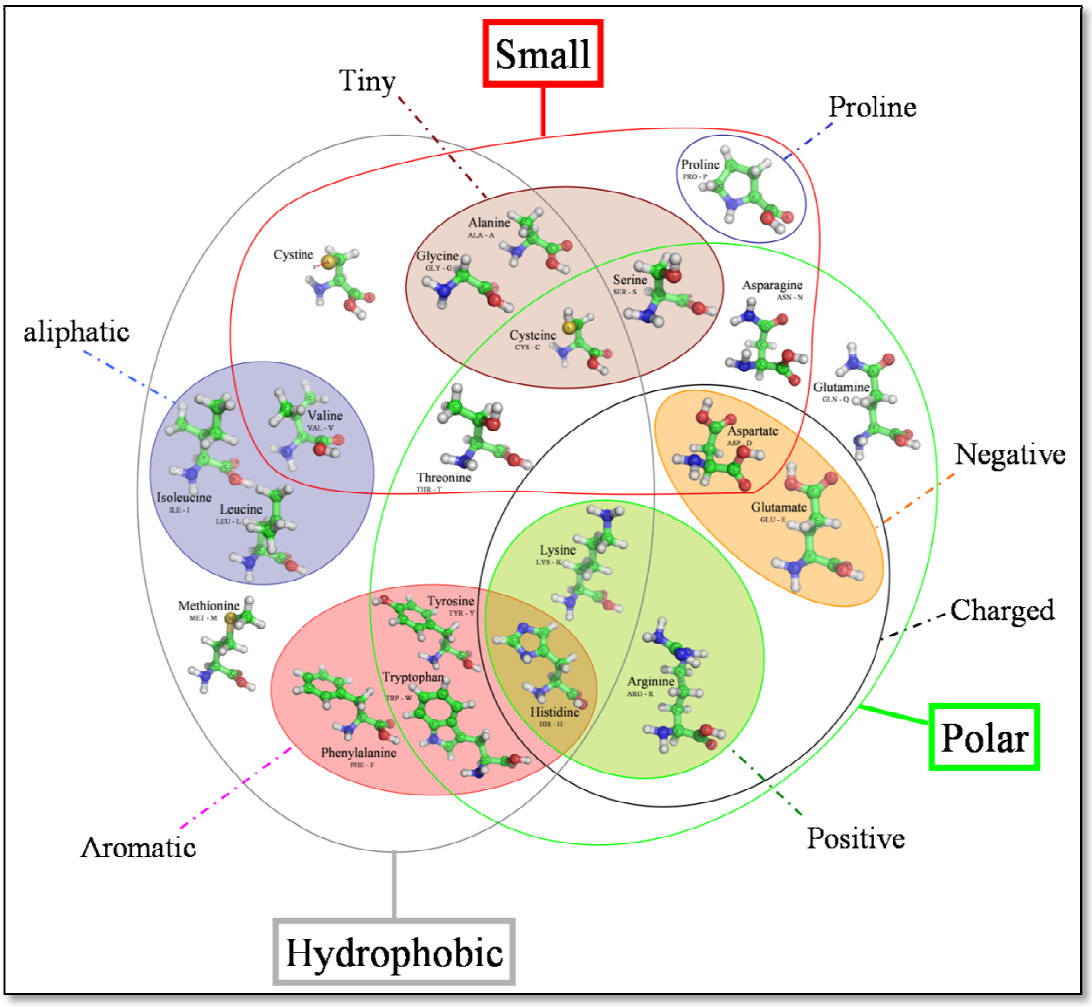
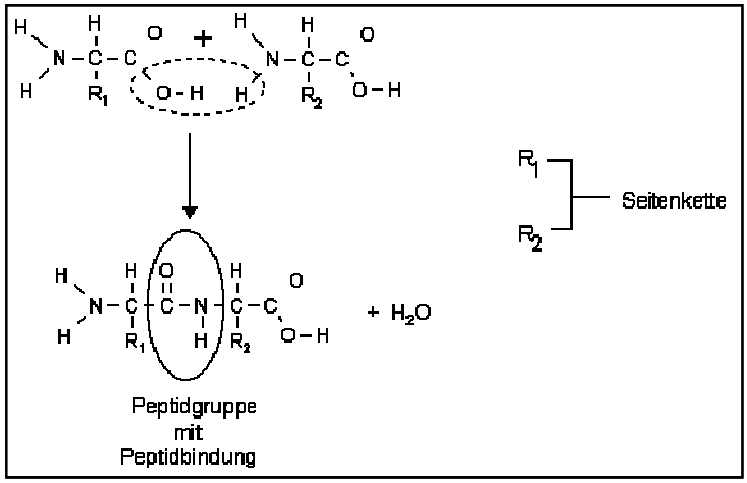


Abbildung 3: Venn-Diagramm der Aminosäuren [19]

Mit Hilfe des molekularen Aufbaus der Restgruppen, ist eine Klassifikation und Charakterisierung der kanonischen Aminosäuren anhand ihrer chemischen Eigenschaften möglich

Bei der Kondensationsreaktion zweier Aminosäuren kommt es zur Bildung einer kovalenten Bindung, der so genannten Peptidbindung, zwischen der Carboxyl-Gruppe der ersten Aminosäure und der Amino-Gruppe der zweiten Aminosäure (Abbildung 4).



; zur
ruppe

Diese entstehende Peptidkette bildet die erste Abstraktionsebene der Proteine: die Primärstruktur. Im Laufe der Zeit hat sich die Darstellung dieser Struktur als Abfolge der Aminosäuren im Einbuchstabencode vom N- zum T-Terminus durchgesetzt. Neben dem Einbuchstabencode ist die Notation im Dreibuchstabencode verbreitet. So ist zum Beispiel die Primärstruktur eines Proteins in den PDB-Dateien der PDB-Webdomain im Dreibuchstabencode abgelegt. Eine Übersicht dieser Notationsweisen ist in Tabelle 1 dargestellt [6].

Tabelle 1: Ein- und Dreibuchstaben-Notation der Aminosäuren [21]

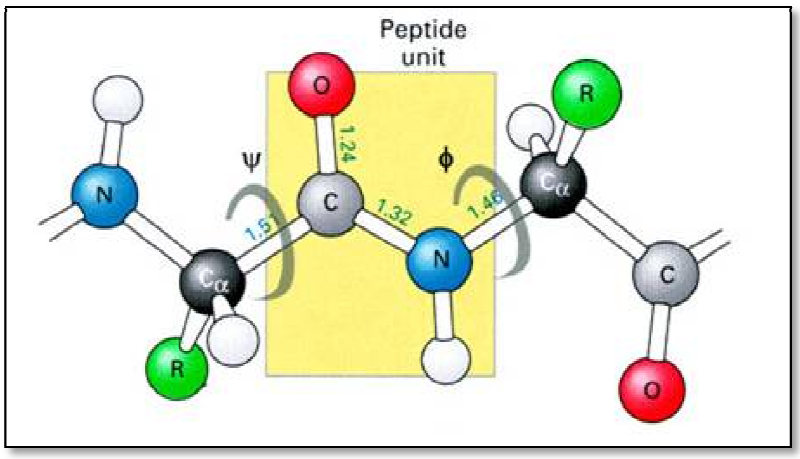
Diese Tabelle gibt einen Überblick über die verschiedenen Bezeichnungsarten. In dieser Arbeit wird überwiegend die Einbuchstaben-Notation verwendet

Name	Dreibuchstabencode	Einbuchstabencode
Alanin	Ala	A
Arginin	Arg	R
Asparagin	Asn	N
Asparaginsaeure	Asp	D
Cystein	Cys	C
Glutamin	Gln	Q
Glutaminsaeure	Glu	E
Glycin	Gly	G
Histidin	His	H
Isoleucin	Ile	I
Leucin	Leu	L
Lysin	Lys	K
Methionin	Met	M
Phenylalanin	Phe	F
Prolin	Pro	P
Serin	Ser	S
Threonin	Thr	T
Tryptophan	Trp	W
Tyrosin	Tyr	Y
Valin	Val	V

Die nächsthöhere Abstraktionsebene ist die Sekundärstruktur. Dabei handelt es sich um regelmäßige Substrukturen der dreidimensionalen Anordnung des Hauptkettenverlaufs (backbone) einer Peptidkette. Wichtig ist zu beachten, dass auf dieser Ebene der Proteinstrukturabstraktion die räumliche Anordnung der Seitenketten nicht mit einbezogen wird [8]. Letztendlich wird die räumliche Anordnung der Peptidkette auf reguläre Faltungsmuster differenziert. Die bekanntesten sind α-Helices, Strands (die aus den β-Sheets der räumlichen Struktur abgeleitet werden), Loops, und Turns. In der vorliegenden Arbeit werden die Sekundärstrukturelemente zu α-Helices, Strands und alle anderen Faltungsmuster zu Random Coils zusammengefasst.

Welches Faltungsmuster in einem backbone-Abschnitt vorliegt, wird durch die intermolekularen Kräfte zwischen zwei oder mehreren Aminosäuren und der daraus folgenden räumlichen Konformation determiniert. Sind die φ- und ψ-Torsionswinkel (siehe Abbildung 5) mehrerer aufeinander folgender Aminosäuren konstant, so bilden

sich helikale Strukturen aus [8], die durch Wasserstoffbrückenbindungen zwischen einer CO-Gruppe einer Residue und der NH-Gruppe der viertnächsten Aminosäure zusätzlich stabilisiert werden. Eine vollständige Drehung erfolgt nach 3,6 Aminosäuren [8].



hreibbar

Ein weiteres Element der Sekundärstruktur ist das β-Faltblatt (engl.: β-Sheet). Dabei handelt es sich um mindestens zwei β-Stränge (engl.: Strands), die durch Wasserstoffbrücken miteinander verbunden sind. Diese bilden sich zwischen den Carboxyl-Gruppen eines Stranges und den Amino-Gruppen eines anderen Strands aus. Auf Grund der Abstraktion der dreidimensionalen Struktur, werden die β-Faltblätter auf die β-Stränge differenziert. Folglich kann man aus der Sekundärstruktur keine Aussage treffen, welche Strands sich zu Faltblättern ausbilden [8, 6].

2.2.2 Tertiär- und Quartärstruktur der Proteine

Durch die räumliche Anordnung der Sekundärstrukturelemente, aller Seitenketten und Atome entsteht, unter dem Einfluss der Wechselwirkungen zwischen den einzelnen Residuen, die Tertiärstruktur. Demnach erhalten Helices, β-Faltblätter und Random Coil-Strukturen räumliche Ausrichtungen. Ein wichtiger Faktor zur Stabilisierung der Tertiärstruktur ist die Bildung der nicht-kovalenten Wasserstoffbrückenbindungen. Weiterhin können die Sulfid enthaltenden Aminosäuren Methionin und Cystein

untereinander die weitaus stabileren Disulfidbindungen ausbilden. In der Mehrzahl aller Fälle verbinden sich sequenzferne Aminosäuren über Disulfide. Dadurch wird die gefaltete Struktur des Proteins zusätzlich stabilisiert. Die Abbildung 6 zeigt zwei durch eine Disulfidbrücke verbundene Cysteine.

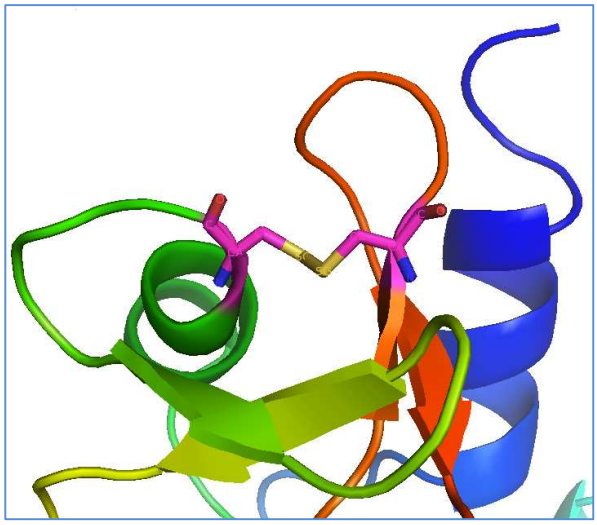


Abbildung 6: Disulfidbrücken-Bindung zweier Cysteine
Die auf die Struktur stabilisierend wirkenden Disulfidbrücken sind selten. Oftmals liegen die verbundenen Cysteine sequenziell sehr weit voneinander entfernt. In diesem Falle beträgt die sequenzielle Distanz 50 Aminosäuren

Im Gegensatz zur Sekundärstruktur ist es in dieser Abstraktionsebene möglich einzelne Strands zu Faltblattstrukturen zusammen zufassen. Dabei zeigt sich hier das Phänomen, das Strands abhängig von ihrer Lage im Protein, parallel bzw. antiparallel zueinander liegen können. So bestehen β-Faltblätter im Inneren eines Proteins meist aus parallel zueinander verlaufenden Strands, während hingegen am Proteinäußeren die Strands antiparallel ausgerichtet sind. Idealisiert betrachtet, ist das Innere des Proteins hydrophob, das Äußere hingegen hydrophil. Das hat zur Folge, das Strands in der Peripherieregion einen charakteristischen Wechsel von hydrophoben und hydrophilen Aminosäuren aufweisen. Während erstere zum Proteininneren ausgerichtet sind, zeigen letztere zum Äußeren hin. Dieser Sachverhalt ist besonders für Secondary Structure Prediction Programme (SSP) interessant.

Proteine, die sich aus mehreren Untereinheiten (Monomeren) zusammensetzen und sich folglich mehrere Tertiärstrukturen räumlich in- bzw. aneinander anordnen, bilden eine

so genannte Quartärstruktur aus. Häufig ist erst das dabei entstehende Di- bzw. Oligomer in der Lage, die entsprechende biologische Funktion auszuführen [8, 22].

2.3 Proteinfaltung

Unter Proteinfaltung versteht man den Übergang vom kettenförmigen, ungefalteten Zustand der Aminosäuren zur nativen funktionsfähigen Konformation. Interessant ist hierbei, dass eine Aminosäuresequenz eines Proteins immer ein und dieselbe charakteristische Tertiärstruktur ausbildet. Die Mechanismen der Proteinfaltung sind bisher weitgehend ungeklärt.

Der Tatsache, dass sich Proteine binnen weniger Mikrosekunden in ihren nativen Zustand falten, steht das rein theoretische kombinatorische Problem der möglichen Konformationsräume gegenüber. Dieser Sachverhalt, der in der Biologie als Levinthal-Paradoxon bekannt ist, wurde in den späten 60er Jahren des letzten Jahrhunderts erstmals von Cyrus Levinthal formuliert [8]. In diesem Paradox wird angenommen, dass ein Protein mit k Aminosäuren bei n möglichen Konformationszuständen der Peptidbindung n^{k-1} mögliche Tertiärstrukturen ausbilden könnte. Bei einem eher kleinen globulären Protein mit einer Länge von 150 Residues und einer rein hypothetisch angenommenen Anzahl von Konformationszuständen von $n = 3$, würde die Anzahl der möglichen Zustände rund $1{,}2 \times 10^{71}$ betragen. Dies wäre physiologisch unmöglich [7].

Betrachtet man eine neu synthetisierte, frei im Cytoplasma liegende Aminosäurekette, lässt sich feststellen, dass, aufgrund der unterschiedlichen hydrophoben Eigenschaften der Residues, bestimmte Bereiche der Proteinsequenz mit der wasserähnlichen Umgebung stärkeren Wechselwirkungen unterliegen und somit mehr Energie besitzen als andere Sequenzabschnitte. Demnach besitzt das ungefaltete Protein ein hohes Maß an freier Energie, wodurch das thermodynamische Prinzip der exergonen Reaktion besagt, dass durch die gegebene Entropie S, das Inertialsystem, in dem Falle das ungefaltete Protein, den Zustand der geringsten freien Energie einnimmt. Ähnlich einem Ball, der auf eine Rampe gelegt wird und sofort nach unten rollt, wir das Protein in die Tertiärstruktur mit der geringsten freien Energie hineingezogen [6].

So nimmt man an, dass sich zuerst kurze Bereiche der Aminosäuresequenz falten und dabei sogenannte Faltungsnuclei bilden. Im Anschluss nehmen diese Faltungsnuclei ihre räumliche Konformation ein. Danach faltet sich der Rest der Struktur um die agglomerierten Nuclei.

Da die physikochemischen Wechselwirkungen einer sehr großen Komplexität unterliegen, ist es nicht möglich die räumliche Struktur, die eine Aminosäuresequenz bei der Proteinfaltung einnimmt, aus den reinen Anfangsbedingungen vorherzusagen [6].

Jedoch gibt es Methoden, um auf die Struktur von unbekannten, im Labor exprimierten oder aus der DNA abgeleiteten Proteinen schließen zu können. Ist nur die Sequenz des Proteins bekannt, können mit Hilfe des *Basic Local Alignment Search Tools* (BLAST) komplette Datenbanken nach identischen oder ähnlichen Sequenzen mit aufgeklärten Strukturen durchsucht werden. Sind solche ähnlichen Proteine in der Datenbank abgelegt, kann man über Homology-Modelling-Server die räumliche Struktur des unbekannten Proteins bestimmen. Dafür benötigen die zugrunde liegenden Algorithmen eine Vorlage (*template*). Als *template* dienen die Daten der PDB der bereits bekannten Proteinstruktur. Ein wichtiger Sachverhalt für diese Arbeit ist, dass die Algorithmen eine Sequenzidentität von mindestens 30% zwischen der Sequenz des *templates* und der Sequenz der zu bestimmenden Proteinstruktur aufweisen müssen [7, 8]. Diese Einschränkung soll durch die Prognostizierung und dem Vergleich von Energieprofilen umgangen werden.

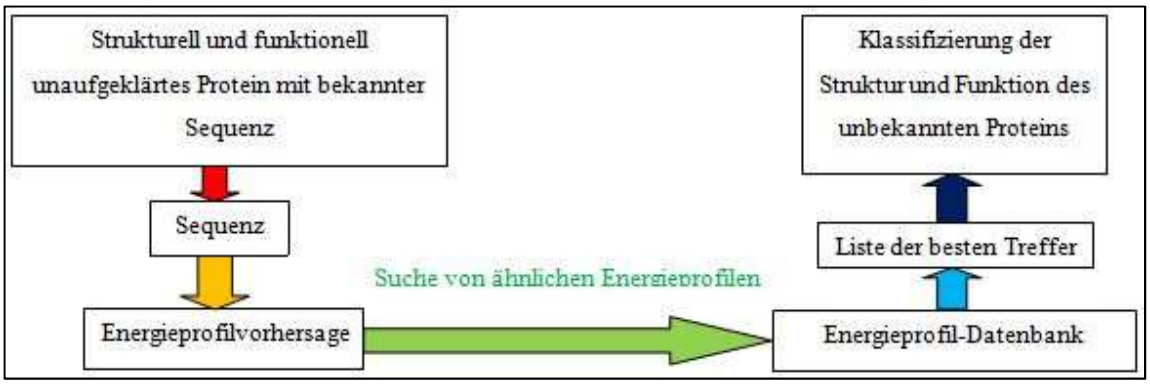


Abbildung 7: Verlauf der Klassifizierung von strukturell unaufgeklärten Proteinen mit Hilfe des vorhergesagten Energieprofils

3. Wechselwirkungen und Energiefunktion

3.1 Intermolekulare Wechselwirkungen

Wie schon in Punkt 2.2.1 beschrieben, werden die physikochemischen Eigenschaften einer Aminosäure durch die Restgruppe determiniert. Folglich sind alle Aminosäuren chemisch exakt anhand ihres Restes charakterisier- und klassifizierbar. Demzufolge sind Rückschlüsse auf chemische Ähnlichkeiten zwischen Aminosäuren möglich. Exemplarisch soll hier die kanonische Aminosäure Isoleucyn heran gezogen werden. Bei diesem Molekül handelt es sich um eine hydrophobe, ungeladene Aminosäure mit einer aliphatischen (linearen) Seitenkette. Die kanonische Aminosäure Arginin besitzt ebenfalls einen aliphatischen Rest, aber unterscheidet sich in den übrigen Merkmalen stark von Isoleucyn. So ist dieses Molekül positiv polarisiert und gehört, unter Normalbedingungen, zur Klasse der hydrophilen Aminosäuren [19]. Demnach sind die Entfernungen zweier Aminosäuren im Venn-Diagramm (siehe Abbildung 3 in Punkt 2.2.1) äquivalent zu chemischen Ähnlichkeit. Je kürzer die Distanz, desto ähnlicher sind die physikochemischen Eigenschaften. Entsprechend wirken auch die Anziehungs- bzw. Abstoßungskräfte. So ist es zum Beispiel eher unwahrscheinlich, dass sich aus der Abfolge von mehreren Arginin- und/oder Lysin-Bausteinen eine stabile Helixstruktur ausbildet. Zwar könnten sich zwischen den Aminosäuren Wasserstoffbrücken ausbilden, jedoch besitzen diese eine zu geringe Bindungsenergie. Die Abstoßungskräfte zwischen diesen identisch geladenen Resten würden die Struktur zum Kollabieren bringen. Demnach gilt: Die räumlichen Zustände, die ein Protein bei der Faltung durchläuft, bzw. die dreidimensionale Struktur, die es danach eingenommen hat, ist zum einen durch die physikochemischen Eigenschaften der Aminosäuren und durch die räumliche Anordnung/Distanz deren Reste zueinander determiniert [2, 28].

Weiterhin spielen die intermolekularen Wechselwirkungen zwischen den Aminosäuren und dem Lösungsmittel eine weitere große Rolle im Kontext der Proteinfaltung bzw. –struktur. Wie im Venn-Diagramm ersichtlich, ist eine der gröbsten aber wesentlichsten Kriterien der Aminosäureklassifikation die Unterteilung nach der Hydrophobizität. Diese chemische Eigenschaft beschreibt das Löslichkeitsverhalten einer Chemikalie in Wasser bzw. in einem wasserähnlichen Lösungsmittel, z.B. Ethanol. Dabei werden alle Stoffe als hydrophob (zu griech. Phóbos = Furcht) bezeichnet, die sich in Wasser nicht

lösen lassen. Umgekehrt bezeichnet man wasserlösliche Substanzen als hydrophil (zu griech. philein = lieben). Ebenso gebräuchlich ist die Bezeichnung lipophil als Synonym für die geringe Löslichkeit in Wasser. Für die Proteinfaltung ist diese Eigenschaft von äußerster Relevanz. In den häufigsten Fällen handelt es sich bei den Medien, in denen sich globuläre Proteine befinden, um Flüssigkeiten mit einem hohen Wasseranteil. Die Theorie des hydrophoben Kollapses besagt, dass bei der Proteinfaltung die Faltungsnuclei[1] durch hydrophobe Aminosäuren entstehen. Dabei bewegen sich diese Residuen vom wasserreichen Lösungsmittel weg und bilden dabei kompakte Aggregationen, die bereits einen geringen Freiheitsgrad besitzen. Die späteren Phasen der Faltung werden ebenso durch die Hydrophobizität mitbestimmt [5]. So sind die klassischen Turnbildner Aspartat, Asparigin, Prolin und Serin allesamt hydrophil [28]. Zudem können regelmäßige Abfolgen von hydrophilen und hydrophoben Aminosäuren, besonders an der Peptidoberfläche, β-Faltblätter erzeugen. Global betrachtet gilt die Aussage, dass sich im Durchschnitt mehr Residuen mit hydrophilen Resten im Proteininneren als auf der Peptidoberfläche befinden [8].

Um den energetischen Zustand einer Aminosäure in einer Peptidkette beschreiben zu können, muss also eine Bildungsvorschrift gefunden werden, die sämtliche Aminosäure-Aminosäure- und Aminosäure-Lösungsmittel-Wechselwirkungen erfasst.

[1] Faltungsnuclei: Kristallisations- bzw. Faltungskeim. Ausgangs- und Angelpunkt der Proteinfaltung.

3.2 Berechnung der Energieprofile von globulären Proteinen

Das physikalische Konzept, dass die Gesamtenergie eines gefalteten globulären Proteins sich aus den durch Wechselwirkungen bedingten Energien der einzelnen Aminosäuren zusammen setzt, ist in der Literatur [4] beschrieben.
Das Grundkonzept zur der Berechnung von Energieprofilen besagt, dass die Gesamtenergie eines Proteins sich aus den durch Wechselwirkungen bedingten Energien der einzelnen Aminosäuren zusammen setzt.

Demzufolge ergibt sich folgende Aussage:

$$E_{Ges} = \sum_{<ij>} e^{*}_{ij} f(r_{ij}) + \sum_{i} e'_{i0} g(i) \qquad (1)$$

Dabei definiert e'_{i0} die Wechselwirkung eines Residue mit dem umgebenden Lösungsmittel. e^{*}_{ij} beschreibt die Interaktion zwischen Aminosäure *i* und Aminosäure *j*, wo bei $f(r_{ij})$ als Funktion des Abstandes zwischen *i* und *j* agiert und die Wechselwirkungen relativiert. *g(i)* definiert den Zustand der Aminosäure.
Um den Abstand r_{ij} ermitteln zu können, muss ein räumlicher Bezugspunkt gewählt werden. Interessant hierfür sind die Positionen der Seitenketten im Raum. Diese lassen sich durch die die Koordinaten der C-β-Atome sehr gut beschreiben. Die Reste zeigen, auf Grund ihrer unterschiedlichen Eigenschaften, zum Teil starke Divergenzen und Konformationen[2] auf. Das C-β -Atom ist als ein Bestandteil der Seitenkette stark an die backbone[3] gebunden und kann als relativ fixiert angesehen werden.
Um die Aminosäure-Lösungsmittel-Wechselwirkungen zu beschreiben, ist es hilfreich die Hydrophobizität der Residuen zu betrachten. Dabei gilt, dass hydrophobe Aminosäuren vom Lösungsmittel in das Innere des Proteins drängen, während hingegen hydrophile Residues Wasserstoffbrückenbindungen mit der Umgebung ausbilden können und keine bzw. kaum Kräfte zum Molekülinneren ausüben. Dieser Zustand wird als *g(i)* definiert und bezeichnet ein Innen/Außen-Kriterium.

[2] Eine Ausnahme bildet Prolin. Durch seine aromatische Seitenkette besitzt es einen weitaus geringeren Freiheitsgrad.
[3] Die backbone beschreibt den Verlauf der Hauptkette des Proteins, wobei die Restgruppen aller Aminosäuren vernachlässigt werden

Als Zentrum des Inneren wurde der Schwerpunkt *c* definiert. Um diesen wurde eine Kugel mit r = 10 Å gesetzt, um den Innen/Außen-Zustandsraum zu definieren.

Dabei ist:

- eine Residue als innen liegend definiert und *g(i)* = 1, wenn gilt

$$|C_\alpha - c| < 5 \vee (C_\alpha - C_\beta)(C_\alpha - c) < 0 \qquad (2)$$

- eine Residue als außen liegend definiert und g(i) = 0, wenn (2) nicht erfüllt ist.

Über ein Set von ca. 2500 globulären Proteinen wurde die Innen/Außen-Häufigkeitsverteilung erstellt.

Tabelle 2: Innen/Außenverteilung der Aminosäuren [4]

Hier ist die statistische Präfernz der Aminosäuren bezüglich eines Innen/Außen-Kriterium aufgeführt. Daraus lässt sich statistisch die freie Energie einer Aminosäure nach der Proteinfaltung ableiten

Aminosäure	innen	außen	Aminosäure	innen	außen
Cys	4582	1016	His	6419	3366
Ile	20370	4141	Gly	16698	14326
Ser	12576	10411	Asp	10001	14327
Gln	7373	7752	Leu	30615	7107
Lys	9285	15193	Arg	11327	10441
Asn	8225	8928	Trp	4001	1193
Pro	9135	9423	Val	23562	6551
Thr	12537	9622	Glu	11165	18091
Phe	13353	2813	Tyr	11228	3529
Ala	22725	11052	Met	7003	1723

Unter Verwendung der Boltzmann-Verteilung lassen sich durch (3) die entsprechenden Lösungsenergien berechnen.

$$e'_{i0} = -k_B T \ln\left(\frac{n_{in}}{n_{out}}\right) \qquad (3)$$

Um die Wechselwirkungen zwischen den Aminosäuren beschreiben zu können, wurde eine Residuen-Residuen-Kontaktstatistik erstellt. Das Konzept ist hier ähnlich des oben erläuterten Ansatzes, mit dem Unterschied, dass sich die Wechselwirkungsenergien zwischen *i* und *j* aus den beobachteten und statistisch erwarteten Kontakten ableiten lassen. Dabei wurden im Datensatz sämtliche Kontakte zwischen den Aminosäuren *i* und *j* ausgezählt und als n_{ij} definiert. Die Konstante $N_{contact}$ beschreibt die Anzahl aller beobachteten Kontakte. Unter Zuhilfenahme der relativen Häufigkeiten p_i und p_j ergibt sich analog zu (3):

$$e_{ij}^{*} = -k_B T \ln\left(\frac{n_{ij}}{N_{contact}\, p_i\, p_j}\right) \qquad (4)$$

Um das reine Paarpotential zu erhalten, müssen nur noch wenige Änderungen vorgenommen werden. Da (3) nur Energiedifferenzen betrachtet, muss entsprechend umgeformt werden:

$$e_{i0} = \left(\frac{1}{i}\alpha_i\right) e_{i0}^{'} \qquad (5)$$

Dabei beschreibt α_i die Anzahl der mit der betrachteten Residue *i* in Kontakt stehenden Aminosäuren *j*, wobei man von einem Kontakt spricht, wenn $r_{ij} < 8$ Å ist. Folglich gilt für das einfache Paarpotential:

$$e_{ij} = e_{i0} + e_{j0} + e_{ij}^{*} \qquad (6)$$

Entsprechend kann man nun nach (6) für jede Aminosäure im Protein die entsprechende Energie berechnen und ein Energieprofil erstellen. Ein Energieprofil ist somit die schematische Auftragung der Energien für jede in der Sequenz vorkommende Aminosäure entlang der Primärstruktur. Es stellt das Ergebnis der Transformation der 3D-Struktur in einem zweidimensionalen Vektor dar.

- Aufgrund der Transformation der chemischen Zusammensetzung und der räumlichen Struktur des Proteins, ist dieser zweidimensionale Vektor für jedes Protein einzigartig. Es handelt sich folglich um einen eindeutigen Fingerprint.

- Demnach kann man das Energieprofil als eine weitere primäre Abstraktionsebene der Proteine verstehen.

Eine Transformation in ein Energieprofil kann mit allen Proteinen vorgenommen werden, von denen eine 3D-Struktur vorhanden ist, und somit die Koordinaten aller Atome vorliegen.
Eine entsprechende Lösung des Algorithmus wurde in Java implementiert. Als Input dient hierfür eine Datei im PDB-Format, wobei die darin abgelegten Koordinaten und Sequenzinformationen zur Berechnung heran gezogen werden.

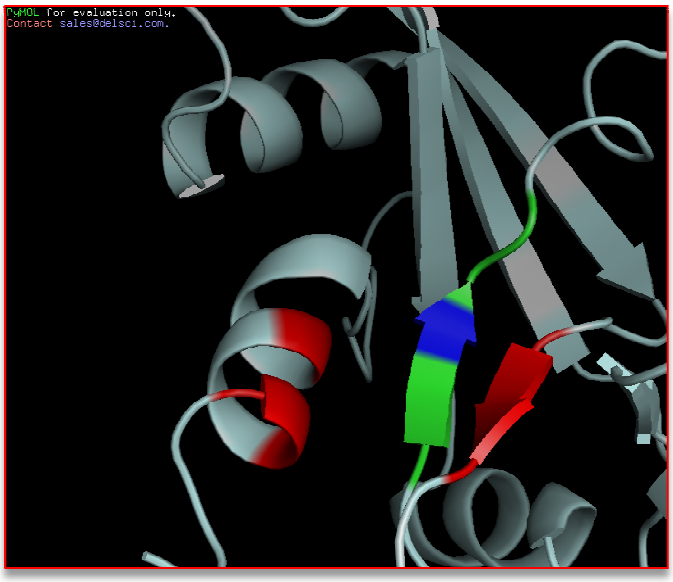
3.3 Ergänzende Informationen zu Energieprofilen

An dieser Stelle soll auf einige wesentliche Aspekte eingegangen werden, die für das Verständnis dieser Arbeit von Relevanz sind.
Zum einen ist dies die physikalische Einheit der Energien. Die üblichste Maßeinheit für E ist das Joule. Im Laufe dieser Arbeit wird jedoch auf die Umrechnung in J[4] oder in $kcal \cdot mol^{-1}$ verzichtet. Dies wird dadurch legitimiert, dass neben der Boltzmann-Konstante k_B auch die Temperatur T als konstant angenommen wird. Folglich sind k_B und T Proportionalitätsfaktoren, die den direkten Zusammenhang zwischen (3,4) und der Energie in Joule beschreiben. Demnach kann auch k_BT in (6) vernachlässigt werden, wobei die physikalische Bedeutung von e_{ij} nicht verloren geht. Folglich handelt es sich bei den in dieser Arbeit erwähnten Energien um einheitslose Werte. Das energetische Spektrum globulärer Proteine reicht von -50 bis +10.
Weiterhin ist das Konzept der Energieberechnung interessant. Diesem Prinzip unterliegt ein rein statisches Modell. Dieses besagt, dass der negative natürliche Logarithmus der Relativität des in Punkt 3.2 beschriebenen Innen-Außen-Kriteriums proportional zur Energie der betrachteten Aminosäure ist. Dadurch greift das Modell, je kanonische Aminosäure, auf nur einen aus diesem Kriterium abgeleiteten Wert zurück. Zur Berechnung der Energie, die eine Aminosäure in einem Protein besitzt, werden nach (6) die $-ln(\frac{in}{out})$ derjenigen Residues addiert, deren r < 8Å zur betrachteten Aminosäure ist. Es handelt sich also bei einer Residue-Energie e_i um die Summe von vordefinierten

[4] Dies geschieht durch die Multiplikation mit k_BT

Werten, wobei die räumliche Struktur des Proteins sowie die sequenzielle Zusammensetzung in der 8Å-Umgebung um *i* erfasst werden und somit die energetische Beschaffenheit der betrachteten Aminosäure beschrieben wird.

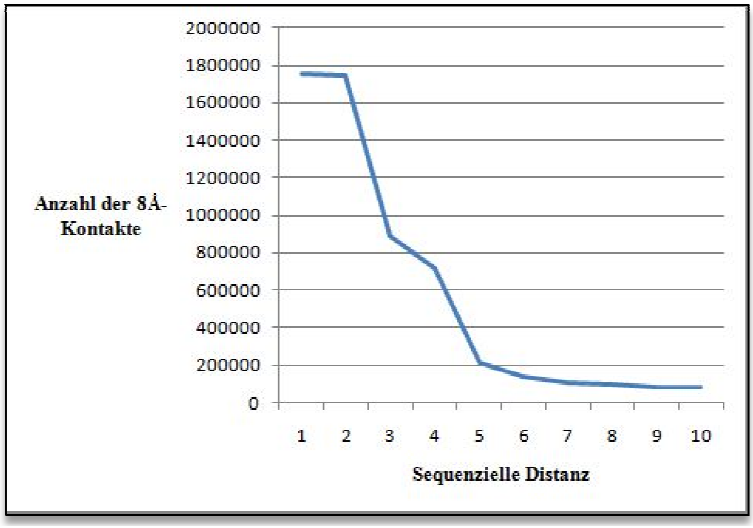
Aus diesem Sachverhalt erschließt sich, dass zum einen sequenziell unmittelbar benachbarte Aminosäure sich gegenseitig energetisch beeinflussen. Zu diesem *lokalen Einfluss* trägt des Weiteren der *globale Einfluss* zu Bildung der Residue-Energie bei. Dieser definiert sich durch jene Aminosäuren, deren $r < 8$Å ist und sich nicht in unmittelbare sequenzieller Nähe zur betrachteten Residue befinden. Die Abbildung 8 zeigt eine solche 8Å-Umgebung exemplarisch am katalytisch aktiven His114 des Angiogenins (PDB-ID: 1B1J). Hierbei ist das His114 blau eingefärbt. Die Residues, die dieses Histidin *lokal* beeinflussen, sind grün koloriert. Entsprechend handelt es sich bei den rot gefärbten Aminosäuren um *global* beeinflussende Residues. Diese befinden sich in einer der drei Helices des Proteins bzw. in einem antiparallelen Strand.



B-ID: 1B1J)

ne Residues (rot)

Des Weiteren wurde untersucht, wie viele Aminosäuren den *lokalen Einfluss* bestimmen. Hierfür wurden alle Aminosäuren und deren 8Å-Kontakte auf sequenzielle Nachbarschaft hin überprüft. Dabei zeigte sich, das im Schnitt $i \pm 3$ sequenziell benachbarte Aminosäuren mit der Residue *i* Wechselwirkungen unterliegen (siehe Abbildung 9).



osäuren

nosäuren in ihrer

ıuenziell entfernt

Die Betrachtung der sequenziellen Distanzen der *global* beeinflussenden Aminosäuren zeigte, dass deren Anzahl und Distanzen starken Schwankungen unterliegen. Dies lässt sich durch die Faltung des Proteins begründen. Eine Residue, die zusätzlich von vielen sequenziell entfernten Aminosäuren umgeben ist, ist folglich fest im Inneren des Proteins eingebettet. Aus der Summation der statischen Werte dieser Aminosäuren ergibt sich für diese Residue eine niedrige Energie. Somit existiert ein Zusammenhang zwischen der Energie und Stabilität einer Aminosäure.

Nähere Erläuterungen, die diesen Zusammenhang bekräftigen, sind im Kapitel 5 aufgeführt.

4. Entwicklung eines Super-Alignments

Als eines der wichtigsten Verfahren der Biologie als beobachtende Wissenschaft ist das Vergleichen von Organismen. Zweck dafür ist es auf phylogenetische Zusammenhänge schließen zu können. Bereits Darwin stütze seine Theorie der Artenentwicklung auf Ähnlichkeitsbetrachtungen der Morphologien verschiedenster Lebewesen, z.B. die Schnabelbeschaffenheit der Galapagosfinken [7]. Dieses Prinzip ist ebenso auf den Nanokosmos der Proteine übertragbar. Aus dem Vergleich zweier Sequenzen können Unterschiede und Ähnlichkeiten aufgedeckt, und somit Rückschlüsse auf die funktionellen, strukturellen und evolutionären Beziehungen gewonnen werden. Mutationen und Veränderungen in Proteinsequenzen erfolgen durch den evolutionären Druck. Dieser verändert den Genpool und damit auch die Proteinsequenzen durch entsprechende Deletionen und Insertionen. Folglich unterliegen Proteine einem ständigen Funktions- und Strukturwandel. Dabei gilt das folgende Paradigma: Eine hohe Sequenzähnlichkeit führt zu einer ähnlichen Struktur und Funktion, während hingegen nahezu identische Funktionen und Strukturen nicht zwangsläufig eine hohe Sequenzähnlichkeit bedeuten [8].

In diesem Kapitel wird eine Alignmentmethode auf Basis des Needleman-Wunsch-Algorithmus vorgestellt, das, neben der reinen Sequenz, ebenso die Informationen der Sekundärstrukturen und die der Energieprofile zweier Proteine zusammenführt und daraus ein einziges paarweises Alignment erzeugt.

4.1 Globales Sequenzalignment nach Needleman-Wunsch

4.1.1 Theoretische Grundlagen

„A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins“- so lautete der Titel des 1969 veröffentlichten Papers, in dem Saul B. Needleman und Christian D. Wunsch ihren Algorithmus zum Erstellen eines globalen Alignments rein informell beschrieben und für die Öffentlichkeit zugänglich machten.

Die Basis für diesen Algorithmus lieferten die beiden Mathematiker Wesley Hamming und Vladimir Levenshtein. Beide befassten sich mit der Problematik, diskrete Beschreibungen und Lösungen bei der Umwandlung zweier Zeichenketten in einen gemeinsamen Konsensus zu finden. Während Hammings Lösung dieses Problems darin bestand, die Anzahl der nötigen Editieroperationen zu minimieren, übertrug Levenshtein die Problematik auf zwei unterschiedlich lange Sequenzen und führte zu jeder Editieroperation eine Gewichtung ein. Mit Hilfe dieser definierten Scoring-Funktionen konnte so der Konsensus bewertet werden. Die Levenshtein´sche Lösung liegt dabei in der Minimierung der Gesamtzahl aller Kosten[5] [8].

Needleman und Wunsch erkannten, dass in der Optimierung des Konsensus zweier Zeichenketten, unter Verwendung von gewichteten Editieroperationen, die Lösung des Problems des paarweisen Aminosäure-Sequenzvergleiches bestand.

Die erste und bis heute genutzte Gewichtung von Editierschritten, berechnete Margaret Dayhoff in den 1970er Jahren. Diese, als PAM-Matrix bekannte Substitutionsmatrix, gilt als eine der ersten Fusionen von theoretischer Informatik, Mathematik und Biologie. Diese und die in den darauf folgenden Jahren entwickelte Substitutionsmatrizen beschreiben die Mutierbarkeit einer Aminosäure oder, genauer gesagt, die logarithmierte Wahrscheinlichkeit, dass sich eine Aminosäure *i* im Laufe der Evolution durch die Aminosäure *j* austauscht [8]. Zur näheren Erläuterung soll hier das Tryptophan als Beispiel heran gezogen werden.

[5] Im Gegensatz zu Levenshtein wiesen Needleman und Wunsch „schlechten“ Editieroperationen (gap-openings) negative Kosten zu, sodass nach N. und W. eine Score-Maximierung und keine Kostenminimierung vorliegt.

In Tabelle 3 sind die Mutierbarkeiten des Tryptophans gegenüber allen 20 kanonischen Aminosäuren aufgeführt[6]. Diese beobachteten Werte gehen einher mit den chemischen Eigenschaften. Betrachtet man das Venn-Diagramm (Abbildung 2) fällt auf, das Tyrosin und Phenylalanin dem Tryptophan physikochemisch sehr ähnlich sind. Alle drei Aminosäuren sind hydrophob und bilden mit einer aromatischen Seitenkette eine sehr kleine exklusive Gruppe unter den kanonischen Aminosäuren. Entsprechend ist es möglich und biologisch nachvollziehbar das sich Y, W und F entsprechend untereinander austauschen können (Mutierbarkeit >0). Anderseits ist es biologisch weniger sinnvoll, dass andere Aminosäuren den Platz von Tryptophan einnehmen können (Mutierbarkeit < 0) [8].

Tabelle 3: Mutierbarkeit des Tryptophans [30]

Unter Mutierbarkeit versteht man die Präferenz einer Aminosäure im Laufe der Evolution durch eine andere Aminosäure ausgetauscht zu werden. Im Falle des Tryptophans ist eine Mutation zu Asparigin (N) eher selten. Diese Substitutionsmatrizen liefern die Gewichtungsfunktion beim Proteinsequenzvergleich.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3

Needleman und Wunsch nutzten diese Substitutionsmatrizen, um entsprechende Scoringfunktionen für Editieroperationen (Matches- bzw. Mismatches, Einführen von Lücken) beim Sequenzvergleich aufzustellen. Die eigentliche bahnbrechende Neuerung dieses Algorithmus ist es, dass, neben den Kosten (Score), auch das optimale Alignment durch ein Traceback-Verfahren ermittelt werden kann. Das löst das durch Levenshtein beschriebene kombinatorische Problem, dass zwischen zwei Sequenzen der Länge N nach (7) eine entsprechend große Zahl S an möglichen Alignments möglich ist [31], wobei aus dieser Lösungsmenge das Alignment mit dem höchsten Score gefunden werden muss.

$$S = \frac{2^{2N}}{\sqrt{\pi N}} \qquad (7)$$

[6] Die aufgeführten Werte sind der BLOSUM62-Matrix entnommen.

Das Verfahren wird als global bezeichnet, weil es die Sequenzen als Ganzes miteinander vergleicht und keine Subalignments erstellt. Das hat allerdings folgende Nachteile: der Algorithmus ist ohne Modifikation nicht auf Sequenzen mit deutlich unterschiedlichen Längen anwendbar und zweitens ist das Verfahren gegenüber kurzen konservierten Bereichen nicht sensitiv genug, sodass diese nicht erkannt werden.

4.1.2 Initialisierung der F- und Pfad-Matrix

Der erste Schritt zur Erstellung des Alignments ist die Initialisierung der Pfad-Matrix (pathMatrix) und der FMatrix. Die FMatrix dient zur Berechnung der einzelnen Scores pro Element. Mit Hilfe der pathMatrix ist es nach der Berechnung möglich, das optimale Alignment zurück zu verfolgen.
Beide Matrizen besitzen die Größe (m+1)(n+1), wobei m bzw. n der Länge der horizontal bzw. vertikal angeordneten Sequenz entspricht. Bei der ersten Spalte sowie Zeile handelt es sich um die φ-Spalte bzw. φ-Zeile [8, 31]. Die Werte in diesen Spalten/Zeilen werden folgendermaßen initialisiert:

$$\begin{aligned} FMatrix(1, j) &= \varepsilon \cdot (j-1) \\ FMatrix(i,1) &= \varepsilon \cdot (i-1) \end{aligned} \tag{8}$$

Dabei definiert ε die vor der Initialisierung festgelegten Kosten für das Einfügen einer Lücke. Eine weitere Möglichkeit der Initialisierung der φ-Spalte/Zeile ist diese mit dem Wert 0 auszufüllen. Diese Methode, auch open-gap oder free-shift genannt, führt dazu, dass Lücken zu Beginn des Alignments nicht mit einberechnet werden und somit der Needleman-Wunsch-Algorithmus gegenüber unterschiedlich langen Sequenzen sensitiviert wird [31].

4.1.3 Berechnung des Alignmentscores

Der zweite Schritt des Algorithmus wird mittels dynamischer Programmierung gelöst. Darunter versteht man, dass Ergebnisse von Teilproblemen zu einem bestimmten Zeitpunkt berechnet werden, sodass diese später, wenn sie benötigt werden, zu Verfügung stehen. D.h. das zur Lösung von Teilproblemen die Resultate vorangegangener Teilprobleme heran gezogen werden [8].
Nachdem die FMatrix und Pfad-Matrix initialisiert wurden, folgt die eigentliche Berechnung des Alignments. Dabei werden die einzelnen Elemente S_{ij} nach der Berechnungsvorschrift (9) ermittelt[8].

$$S_{ij} := \max\left\{S_{i-1,j} + s(a_i, \varepsilon), S_{i-1,j-1} + s(a_i, b_j), S_{i,j-1} + s(\varepsilon, b_j)\right\} \quad (9)$$

ε entspricht dem Score zum Einführen einer Lücke, während $s(a_i,b_j)$ dem Wert der Substitutionsmatrix an der Stelle (a , b) entspricht.
Die Berechnung beginnt dabei an der Stelle (2,2) und arbeitet sich so Element für Element zeilenweise durch die Matrix, bis das Element (m + 1, n + 1) berechnet wurde. Der Wert dieses Elementes entspricht dem Score des globalen Alignments. Wichtig für das Traceback-Verfahren ist es, das das entsprechende Element, das den Wert S_{ij} definiert, in der Pfad-Matrix in (i, j) abgelegt wird.
Als Pseudo-Code lässt sich das Verfahren wie folgt darstellen [8]:

```
Für i = 2, 3,…n führe aus
        Für j = 2, 3,…m führe aus
                mLEFT   ← FMatrix[j-1, i] + gap-cost
                mDIAG   ← FMatrix[j-1, i-1] + s(a_i,b_j)
                mTOP    ← FMatrix[j, i-1] + gap-cost

                FMatrix[i, j] ← max(mLEFT, mDIAG, mTOP)

                Wenn FMatrix[i, j] == mDIAG
                        pathMatrix[i, j] ← "D"
                oder  FMatrix[i, j]  == mLEFT
                        pathMatrix[i, j] ← "-"
                ansonsten
                        pathMatrix[i, j] ← "|"
```

4.1.4 Traceback-Verfahren

Das Backtracing folgt im Anschluss der Alignmentberechnung. Dafür macht man sich die Elemente der Pfad-Matrix zu nutze. Der Einsprungspunkt des Verfahrens ist das Element (m, n). An dieser Stelle werden sämtliche benötigten Strings und Zahlenwerte initialisiert und im Anschluss einer Methode übergeben, die sich solange rekursiv aufruft, bis das Element (1, 1) erreicht ist [8]. Dabei werden Strings zur Erzeugung des Alignments aufrufweise erweitert.

Als Pseudo-Code lässt sich der Algorithmus wie folgt darstellen:

```
initTraceback()

        m ← Länge der Sequenz A + 1
        n  ← Länge der Sequenz B + 1
        String aligntA ← „“
        String aligntB ← „“

traceback(m, n, aligntA, aligntB)
```

```
traceback(m, n, aligntA, aligntB)
        Wenn m == n == 1
                output(aligntA, aligntB)
               Ende

        Wenn pathMatrix[m, n] == “D”
                aligntA ← char in Sequenz A an der Stelle m + aligntA
                aligntB ← char in Sequenz B an der Stelle n + aligntB
                m ← m -1
                n  ← n – 1

        oder  pathMatrix[m, n] == “-”
                aligntA ← char in Sequenz A an der Stelle m + aligntA
                aligntB ←  „-“ + aligntB
                m ← m -1
                n  ← n

         ansonsten
                aligntA ←  „-“ + aligntA
                aligntB ←  char in Sequenz B an der Stelle n + aligntB
                m ← m
                n  ← n -1

traceback(m, n, aligntA, aligntB)
```

4.2 Struktur-Struktur- und Energie-Scoringfunktionen

In diesem Abschnitt sollen die verwendeten Scoringfunktionen für die verschiedenen Informationen erläutert werden.
Für den Vergleich auf Primärstrukturebene wurde die BLOSUM62-Matrix gewählt. Diese gehört, neben der PAM250-Substitutionsmatrix, zu den am meisten verwendeten Matrizen und ist als Einstellungsparameter auf jeder Alignment-Webdomain zu finden.
Für das Einführen einer Lücke wurde eine gap-cost ε von -40 definiert, wobei gap-extensions nicht berücksichtigt wurden.
Auf Sekundärstruktureben wurde ein einfaches gleich/ungleich-Kriterium eingeführt, wobei gilt:

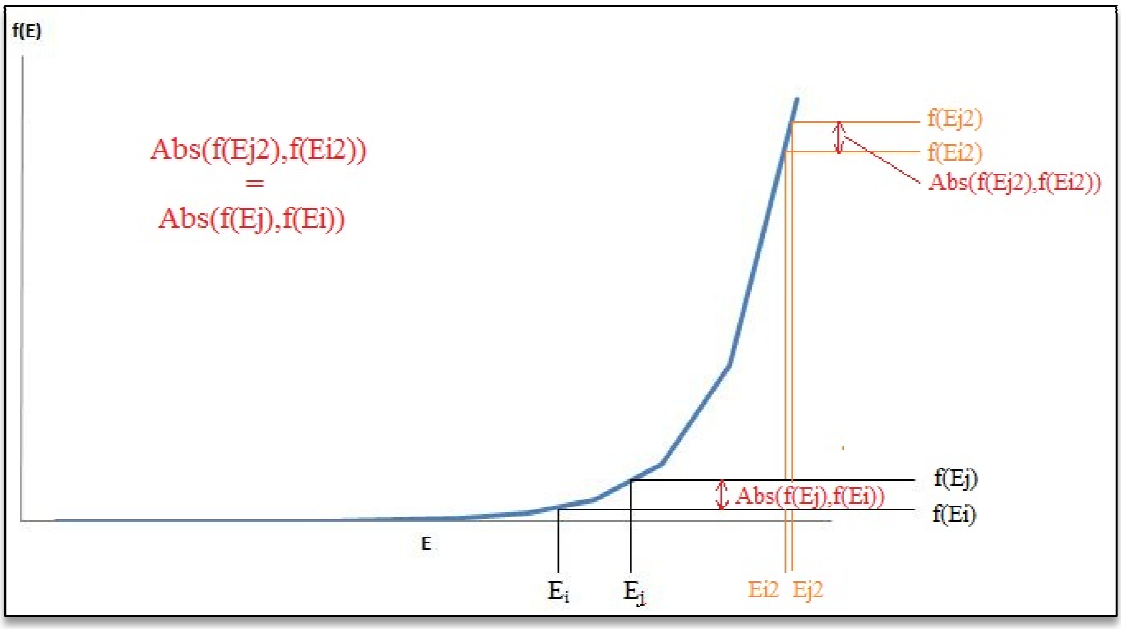
$$S_{S-S} = \begin{cases} -5 : S_{a_i} \neq S_{b_j} \\ +5 : S_{a_i} = S_{b_j} \end{cases} \qquad (10)$$

Das Sekundärstruktur-Scoring $S_{S\text{-}S}$ von 5 bzw. -5 wurde nach der Testphase des Algorithmus ermittelt.
Für die Bewertung der Energien musste eine Vergleichsmöglichkeit gefunden werden, die die Energiewertepaare in ein Scoring-Intervall von -10 bis +12 überführt, wobei das Scoring ganzzahlige Werte annehmen sollte.
Die Grundidee bestand darin, die Energien anhand ihrer Ähnlichkeit zu betrachten. Allerdings ist es nicht möglich den Betrag des Wertepaares zum Scoring heran zu ziehen. Vielmehr sollten die einzelnen Energien so transformiert werden, dass die Beträge der daraus resultierenden Werte sich für das Scoring eigneten.
Um dieses Problem näher zu erläutern, soll folgendes Denkmodell dienen: Eine Chemikalie, z.B. Wasser, befindet sich in seinem festen Aggregatzustand, in diesem Fall handelt es sich um Eis. Die Wassermoleküle sind zu einem stabilen Kristallgitter angeordnet. Das Maß der Schwingung der Moleküle in allen drei Freiheitsgraden ist äquivalent zur Energie, die das System besitzt. Nimmt man zum Beispiel an, dass das Eis bei einer Temperatur von -100°C vorliegt und nun das System auf -90°C erhitzt wird, so kann man beim Vergleich der beiden Zustände feststellen, dass sich das Maß an Molekülschwingung im Kristallgitter nur unwesentlich geändert hat. Betrachtet man dagegen 80°C heißes Wasser und möchte dieses ebenfalls um 10°C erhitzen, wird man

feststellen, dass man dafür um einiges mehr Energie benötigt, als bei -100°C kalten Wasser. Die Molekülbewegung erhöht sich dabei um ein vielfaches [32]. Demnach lassen sich Systeme mit unterschiedlichen Energien nicht eindeutig vergleichen.
Um dieses Problem zu lösen wurde folgendes angenommen:

- zwei Residuen mit unterschiedlich niedrigen (stabilen) Energien E_i und E_j und deren Betrag *abs(Ei-Ej)* sind sich in ihren Wechselwirkungen ähnlicher als zwei hohe (instabile) Energien mit dem selben Betrag.

Folgende Abbildung soll diesen Sachverhalt verdeutlichen:



tialfunktion zu Grunde.
eren, dass zwei Energien

Hier zeigt sich, dass zwei instabile Aminosäuren energetisch viel ähnlicher sein müssen als stabile Residuen, damit der gleiche Betrag und daraus der gleiche Score hervorgeht. Der Transformationsfunktion liegt somit eine Exponentialfunktion der Form $f(x) = e^x$ zugrunde.

Die Untersuchung der Streuung der Energien zeigte, dass diese nicht normalverteilt vorliegen. Dieser Sachverhalt bestätigt, dass zwei Energien nicht über deren Betrag bewertet werden können.

Um dieses Problem zu umgehen, musste ein Transformationsraum $f_T(E)$ gefunden werden, der das Scoring zweier Energien über den Betrag der transformierten Werte ermöglicht. Ein Ansatz zur Lösung dieses Punktes ist die Bildung von gleichmächtigen Intervallen innerhalb des Energiespektrums. Dadurch wird die Verteilung der Energien in eine transformierte Normalverteilung überführt.

Um diese in ihrer Mächtigkeit identischen Intervalle zu identifizieren, wurden alle Energien des Datensatzes sortiert und anschließend die Energiewerte an den durch (11) definierten Indizes festgehalten.

$$Index = \left(\frac{22000 \cdot n}{19} \right) \rightarrow (1,...,n,...,18) \in \mathrm{N} \qquad (11)$$

Diese Energiewerte definieren demnach die Grenzen für 18 energetisch gleichmächtige Intervalle[7]. Diese Bereiche wurden nun zur Transformation der Energien und der anschließenden Bewertung herangezogen. Soll nun der Energie-Score S_E zweier Energien berechnet werden, wird jede Energie seinem entsprechenden Intervall im Spektrum zugeordnet. Der Betrag dieser Intervalle ist nun ausschlaggebend für S_E (12), wobei dieser durch eine experimentell ermittelte Betrag-Score-Tabelle fest definiert ist (siehe Tabelle 4).

$$Abs = \left| f_T(E_i) - f_T(E_j) \right| \qquad (12)$$

[7] Die Menge von 18 Intervallen wurde nach der Testphase mit verschiedenen Intervall-Anzahlen fest gelegt

Tabelle 4: Scoring-Tabelle für Energie-Intervall-Beträge

Zur Bestimmung des Scores zweier Energien, wird jedem Wert das entsprechende Intervall im Energiespektrum zugewiesen. Diese Transformation führt eine Normalverteilung ein, wodurch die Beträge der Intervalle zur Bewertung der beiden Energien herangezogen werden können.

Abs	S_E
0	+12
1	+8
2	+6
3	+2
4	-4
5	-8
>5	-10

4.3 Implementierung

Das Hauptaugenmerk in diesem Kapitel liegt auf der processor-Methode der Implementierung. Diese führt die aus den PDB-Dateien gelesenen Sequenzen und Sekundärstrukturen sowie die berechneten Energieprofile zusammen und berechnet daraus den Score des paarweisen Super-Alignments.

Es muss auch gesagt werden, dass bei der Formatierung der FMatrix die, unter 4.1.2 beschriebene, open-gap-Modifikation verwendet wurde.

Zur Implementierung des Algorithmus bot sich Java als Programmiersprache an. Einer der großen Vorteile gegenüber anderen systemnäheren Sprachen, wie z.B. C++ ist es, das Java, neben seiner Plattformunabhängigkeit, auf komplizierte Steuerung von Array-Zeigern verzichtet. Um auf Elemente in einem Array zugreifen zu können, muss man nur die nötigen Array-Schlüssel angeben. Das ergab den Vorteil, dass die entsprechenden Arrays für Sequenz, Sekundärstruktur und Energie indexgleich abgelegt werden konnten. Beim Zugriff auf die Aminosäure in der Sequenz an der Stelle *i* kann man, ohne *i* ändern zu müssen, auf dessen Energie und Sekundärstruktur zugreifen. Diese Datenstruktur kann man sich als drei gleich lange, parallel nebeneinander liegende ein-dimensionale Arrays vorstellen.

Während das Scoring der Primärstrukturen nach dem unter 4.1.1 beschriebenen Verfahren berechnet wird, werden die Scores der Sekundärstrukturen und Energien durch die in 4.2 beschriebene Vorschrift ermittelt.

- Die erhaltenen Einzelscores werden addiert und ergeben den Gesamtscore des Elementes der FMatrix.

Durch einfache Addition können sich die Scorings der unterschiedlichen Ebenen aufheben. So kann ein negativer Score auf Primärstruktur-Ebene durch eine ähnliche Energie oder ähnliche Sekundärstrukturstruktur ausgeglichen werden. Entsprechend werden Missmatches in allen Abstraktionsebenen hart bestraft.
Der Gesamtscore setzt sich grob aus folgenden Anteilen zusammen:

- 25% aus Primärstruktur-Score
- 50% aus Energie-Energie-Score
- 25% aus Sekundärstruktur-Score

Das modifizierte Verfahren lässt sich im Pseudo-Code folgendermaßen darstellen:

```
Für i = 2, 3,… n führe aus
        Für j = 2, 3,… m führe aus

            mTOP   ← FMatrix[i-1, j] + gap-cost
            mDIAG← FMatrix[i-1, j-1] + S_S-S(S_ai, S_bj) + S_E(E_ai,E_bj) + S(a_i,a_j)
            mLEFT ← FMatrix[i, j-1] + gap-cost

            FMatrix[i, j] ← max(mLEFT, mDIAG, mTOP)

            Wenn FMatrix[i, j] == mDIAG
                    pathMatrix[i, j] ← "D"
            oder   FMatrix[i, j]  == mLEFT
                    pathMatrix[i, j] ← "-"
            ansonsten
                    pathMatrix[i, j] ← "|"
```

4.4 Score-Modifikation

In den ersten Testläufen zeigte sich, dass die Scores selbst bei ungünstigen Alignments schnell sehr groß wurden. Sehr lange Alignments mit vielen großen und kleinen Lücken erzeugten gleiche bzw. sogar höhere Scores als kurze Alignments mit signifikant-identischen Bereichen. Dafür wurde der errechnete Score mit der Länge des Alignments relativiert. Dieser neue Score S_L wird wie folgt berechnet:

$$S_L = \frac{100 \cdot S}{L} \qquad (13)$$

Dabei entspricht S dem Score des Alignments und L ergibt sich aus dessen Länge.
Ergibt z.B. ein Alignment A, mit einer Länge von 200 Zeichen, einen Score von 800, so ist dessen S_L von 400 doppelt so groß wie der eines Alignments mit einem Score von 1000 und einer Länge von 500 Zeichen. Der S_L stellt demnach die Relativität der Alignmentlänge und den durchschnittlichen Score pro Editieroperation dar. Dadurch wird ein grober Überblick über die Signifikanz und Güte des Alignments ermöglicht.

4.5 Evaluierung des Verfahrens und Ergebnisdiskussion

4.5.1 All-against-all-Evaluierung

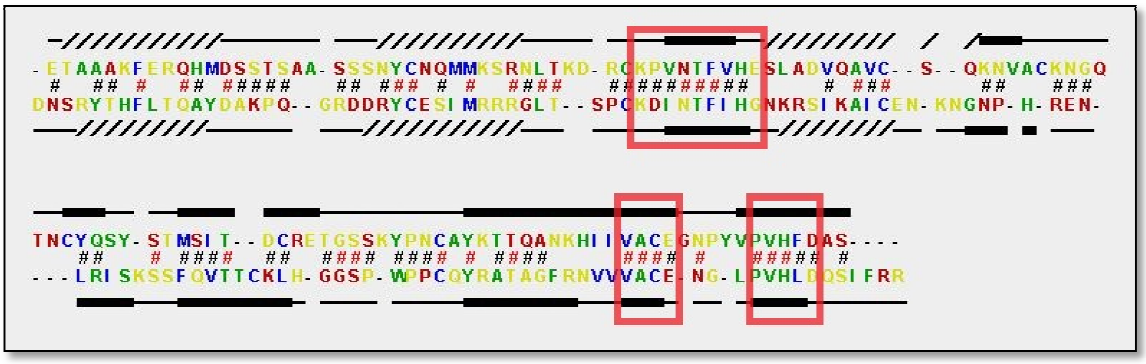
Bei dieser Evaluierung wurden mit Hilfe des Algorithmus alle Proteine zueinander paarweise alignt. Das Ziel bestand darin, Homologien zwischen Proteinen zu finden, die auf der Ebene der Sequenz nicht hätten aufgedeckt werden können.
Hierfür wurde das Programm so modifiziert, das es über einen Datensatz von 118 Proteinen mit Hilfe von Iterationen paarweise Super-Alignments durchführt, wobei Wiederholungen ausgeschlossen wurden. Zudem sollte der Output von nicht signifikanten Alignments unterbunden werden. Hierfür wurde ein Cut-off-Score von 400 festgelegt. Unterschreitet der Score eines Super-Alignments diesen Wert, wird die weitere Prozessierung abgebrochen. Der Output signifikanter Ergebnisse wurde in einer Textdatei festgehalten.

Nach dem Durchlauf von 6642 Alignments blieben ca. 10-15 interessante Ergebnisse übrig.

4.5.2 Exemplarisches Ergebnis und Diskussion

In diesem Abschnitt möchte ich auf ein Ergebnis des Testlaufs eingehen. Dabei handelt es sich um das Alignment eines Proteins der Ribonuclease A-Familie (PDB-ID: 1FS3) und einem Vertreter der Angiogenine (PDB-ID: 1B1J). Das Resultat schien bemerkenswert zu sein, da es in der frühen Entwicklungsphase des Algorithmus bisher nie zum Vorschein gekommen war[8].

Abbildung 11 zeigt das Super-Alignment der beiden Proteine.



struktur-Elemente
erechten Geraden
erdeutlichen Coil-
enzalignment, die
Energie (Näheres
nosäuren.

[8] Zu diesem Zeitpunkt war das Alignen nur auf Sequenzebene möglich

An dieser Stelle sollen einige erklärende Punkte helfen, den graphischen Output des Super-Alignments zu verstehen:

- Die erste Zeile zeigt die alignte Sekundärstruktur des ersten Proteins (in dem Fall 1FS3)
- Die Sekundärstrukturelemente sind optisch folgendermaßen gegliedert:
    - Schräge Linien entsprechen Residues, die in Helices vorliegen
    - Verdickte waagerechte Linien entsprechen Residues, die in Strands vorliegen
    - Schmale waagerechte Linien entsprechen Residues, die in Coils vorliegen
- Die zweite Zeile zeigt nun die alignte Sequenz des ersten Proteins
- Die Färbung der Sequenz erfolgt nach dem energetischen Quantil[9]
- Die Darstellung der alignten Sequenz (Zeile 4) und Sekundärstruktur (Zeile 5) des zweiten Proteins erfolgt analog zu den obigen Punkten
- Lücken in Sequenzen und Sekundärstrukturen werden durch Bindestriche bzw. Leerzeichen dargestellt
- Zur Darstellung der Similarity oder Identity wurde folgendes festgelegt:
    - Wenn zwei alignte Aminosäuren in zwei von drei Charakteristika übereinstimmen, wird diese Position als *total similarity* bezeichnet und durch eine schwarze Raute gekennzeichnet
    - Wenn zwei alignte Aminosäuren in allen drei Charakteristika übereinstimmen, wird diese Position als *total identity* bezeichnet und durch eine rote Raute gekennzeichnet

An diesem Alignment sind verschiedenste Dinge auffällig. Zum einen liegt der S_L mit einem Wert von 850 in einem interessanten Bereich. Während Alignments von nahezu identischen Strukturen, Energien und Sequenzen Werte um 2000 erzielten und nicht signifikante Alignments mit kurzen, eher zufälligen identischen Bereichen maximal einen Score von 400 erreichten, lag dieser Wert über der twilight-zone [10] [11]. Mit nur 39

[9] Nähere Erläuterungenen sind unter Punkt 5.2 zu finden

[10] Der Wertebereich des Scorings in dem anhand des Alignments nicht eindeutig festgestellt werden kann, ob dieses signifikant ist oder nicht

[11] Dieser Bereich liegt zwischen 400 < SL < 600

identischen Residues, was einen Prozentsatz von ca. 30% ausmacht, liegt die Ähnlichkeit dieser beiden Proteine nicht auf der Ebene der Primärstruktur. Das bestätigte auch das Sequenz-Alignment, das mittels ClustalW2 des EBI durchgeführt wurde (Abbildung 12). Der Score dieses globalen Alignments betrug nur 29.

```
1FS3_A|PDBID|CHAIN|SEQUENCE      -KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHE 49
1B1J_A|PDBID|CHAIN|SEQUENCE      QDNSRYTHFLTQAYDAKPQGRDD-RYCESIMRRRGLTSP-CKDINTFIHG 48
                                  .::  ::*  *  *:.... .. .**:.:*: *.**.  ** :***:*

1FS3_A|PDBID|CHAIN|SEQUENCE      SLADVQAVCSQKNVACKNGQTNCYQSYSTMSITDCRETGSSKYPNCAYKT 99
1B1J_A|PDBID|CHAIN|SEQUENCE      NKRSIKAICENKNGNPHR--ENLRISKSSFQVTTCKLHGGSPWPPCQYRA 96
                                 .  .::*:*.:**   :.   *   * *::.:* *:  *.* :* * *::

1FS3_A|PDBID|CHAIN|SEQUENCE      TQANKHIIVACEGNPYVPVHFDASV---- 124
1B1J_A|PDBID|CHAIN|SEQUENCE      TAGFRNVVVACENG--LPVHLDQSIFRRP 123
                                 * . ::::****..  :***:* *:
```

oteine nicht erkennbar

Bei der Recherche auf der Web-Domain der PDB zeigte sich, dass es sich bei dem Protein 1FS3 um eine in Rinderpankreas vorkommende Ribonuclease handelt. 1B1J ist hingegen ein Angiogenin des Menschen. Beide Proteine sind aktiv an der chemischen Reaktion der Transphosphorilierung beteiligt, wobei 1FS3 katalytisch cytotoxische RNA an ihren Phosphatgruppen spaltet. 1B1J dagegen katalysiert als Phosphatransferase bei der Gewebs- und Kapillarneubildung (Angiogenese) [12]. Dabei ist seine enzymatische Aktivität weitaus geringer als die des 1FS3 [11, 10].

Die Abbildung 13 zeigt eindrucksvoll, wie ähnlich sich die Proteine tatsächlich sind. In der linken Hälfte ist 1FS3 abgebildet, entsprechend 1B1J in der rechten.

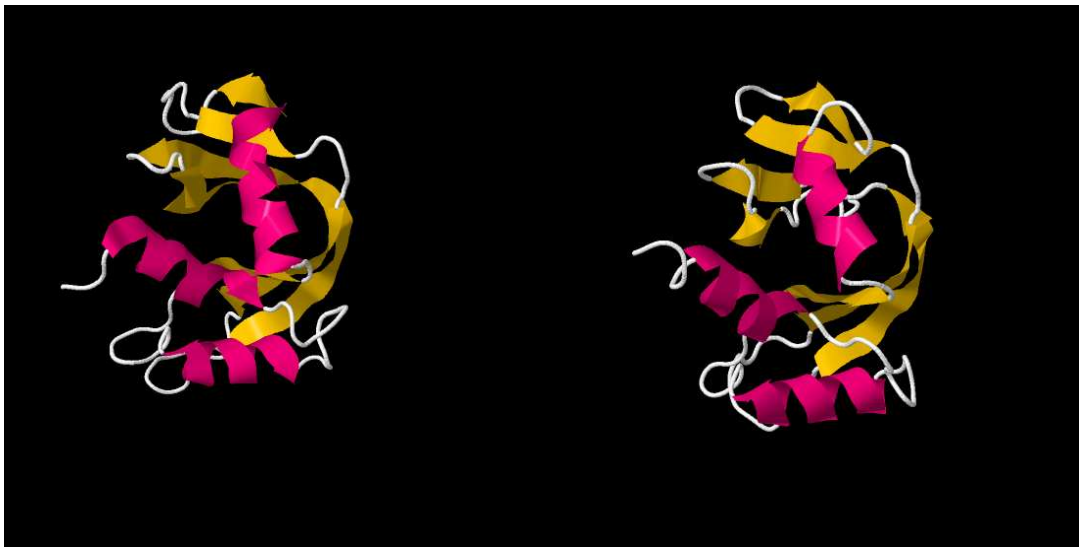


Abbildung 13: Die beiden Proteinstrukturen im direkten Vergleich

Diese Abbildung zeigt eindrucksvoll die strukturelle Ähnlichkeit der beiden Proteine (links: die Ribonuclease 1FS3, rechts: das Angiogenin 1B1J)

Beim Vergleich der beiden Energie-Profile (Ausschnitte der beiden Profile sind in Abbildung 14 dargestellt) zeigt sich jedoch, dass die Grundzüge im Profilverlauf sehr ähnlich sind. Dies lässt den Schluss zu, dass beide Proteine in einen nahezu identischen strukturellen Aufbau besitzen. Dies wird durch das räumliche Alignment (Abbildung 15) bestätigt.

Durch das auf Energieprofilen basierende Alignmentverfahren konnte diese Homologie eindeutig nachgewiesen werden.

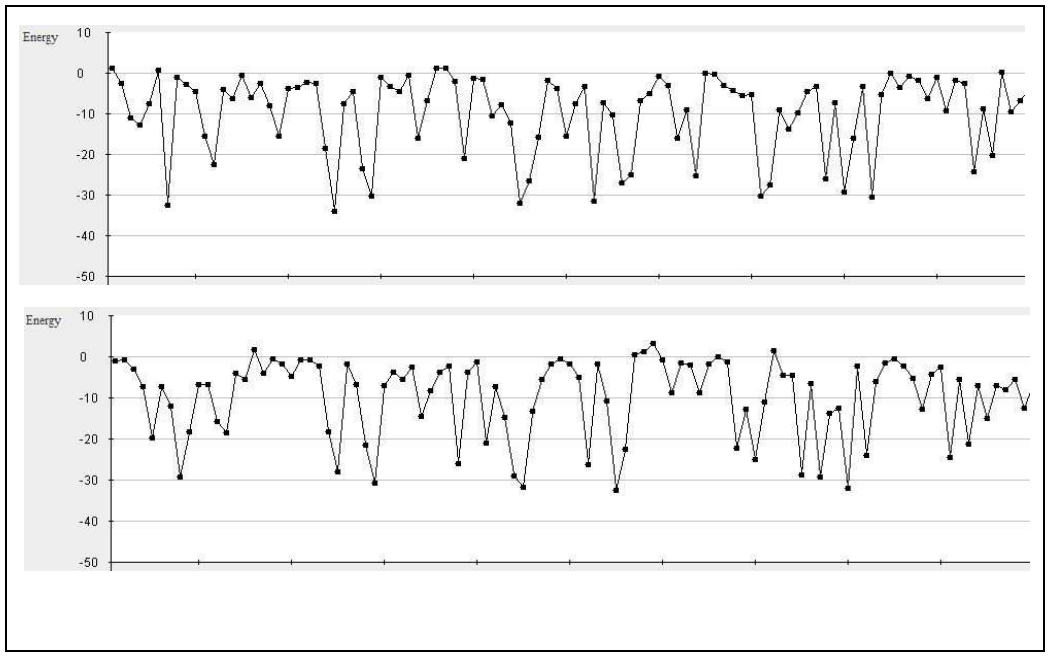


Abbildung 14: Die Energieprofile der Ribonuclease (oben) und des Angiogenins (unten) im Vergleich
Beim Betrachten der Energieprofile sind die Ähnlichkeiten im Profilverlauf auffällig. Daran lassen sich die strukturellen Gemeinsamkeiten beider Proteine ableiten.

An dieser Stelle möchte ich auf die in Abbildung 15 hervorgehobenen Bereiche eingehen. Diese sind von daher interessanter, als die übrigen Regionen, da darin mehrere aufeinander folgende (größtenteils) identische Residuen vorkommen, die zudem in ihrer Energie und Struktur ähnlich bzw. gleich sind. Um diese Stellen in der 3D-Struktur hervorzuheben, wurden beide Proteine in den PyMol-Viewer geladen. Im Anschluss wurden beide Proteine räumlich alignt und die in Abbildung 15 markierten Bereiche farblich hervorgehoben. Bei der weiß gefärbten Struktur handelt es sich um das 1FS3. 1B1J ist entsprechend rot markiert. Die drei Regionen sind orange (1B1J) bzw. grün (1FS3) eingefärbt. Um einen besseren Überblick zu verschaffen, wurde diese Struktur einmal von vorn und im Anschluss noch einmal von der Seite dargestellt.

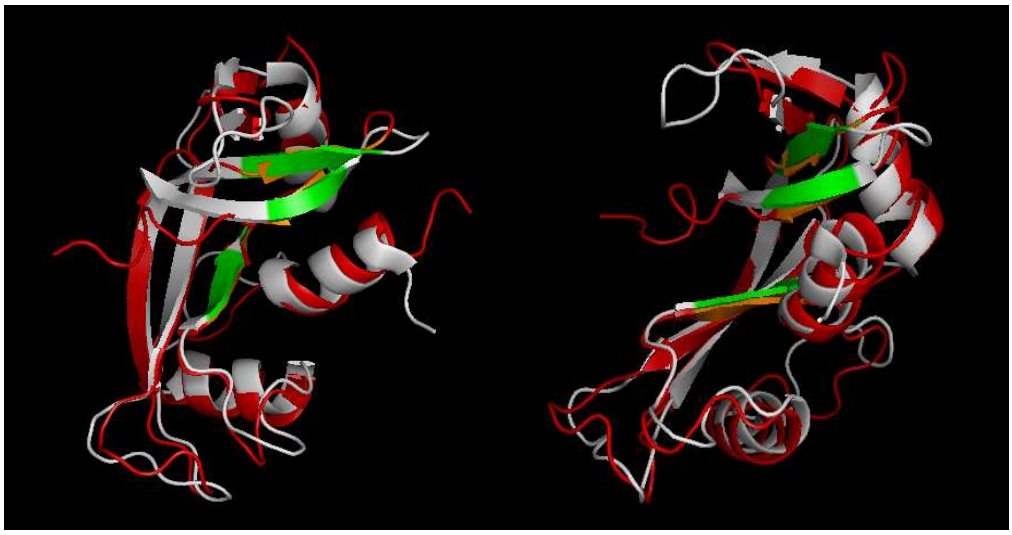


Abbildung 15: Strukturelles Alignment der beiden Proteine

Das Alignment beider Strukturen (weiss: 1FS3, rot: 1B1J) zeigt den hohen Grad an räumlicher Identität. Bis auf die Coil-Strukturen, sind alle weiteren Elemente in ihrer Konformation gleich angeordnet. Die substratbindenden und katalytisch aktiven Aminosäuren wurden zudem farblich hervorgehoben und unterstreichen die funktionelle Gemeinsamkeit beider Proteine anhand ihrer identischen räumlichen Lage.

Hier lässt sich gut erkennen, wie stark sich die Strukturen und markierten Bereiche decken. Bis auf die Coil-Strukturen, sind alle restlichen Tertiärstrukturelemente, räumlich identisch angeordnet. Bei den grün und gelb markierten Regionen handelt es sich um die aktiven Zentren der Proteine. Die Strukturen formen ein sattelförmiges Gebilde. Die dadurch entstehende Vertiefung entspricht der Bindungstasche des Proteins, in der die darin enthaltenen gelb-grün-markierten Bereiche liegen.

Es lässt sich also abschließend festhalten, dass mit Hilfe des Super-Alignments eine Homolgie gefunden wurde, die nicht auf der reinen Sequenzebene hätte identifiziert werden können. Die starke Konservierung der kurzen Bereiche und die gute Übereinstimmung von Sekundärstruktur und Energie bestätigten den evolutionären und funktionellen Zusammenhang beider Proteine. Zwar ist der ermittelte S_L signifikant, zeigt aber auch auf, dass die Enzyme in ihrer Struktur und Funktionalität leicht divergent sind. So besitzt das menschliche Angiogenin 1B1J eine weitaus geringere Enzymaktivität. Ein Grund dafür könnte sein, dass 1B1J in der Konformation der katalytisch aktiven Seitenketten leichte Abweichungen zeigt. Betrachtet man die van-der-Waals-Räume der Aminosäuren, zeigt sich, dass durch diese leichten Änderungen

die Residues des aktiven Zentrums viel Tiefer in der Bindungstasche liegen (siehe Abbildung 16). Dadurch wird die Substratzugänglichkeit verringert, was eine Verminderung der Enzymaktivität zu Folge hat.

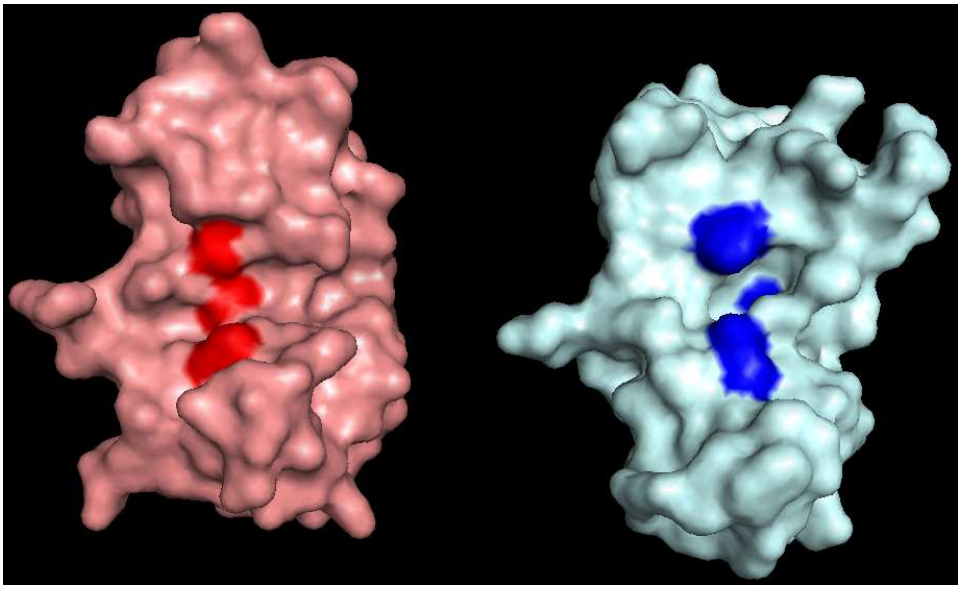


Abbildung 16: van-der-Waals-Oberflächen der beiden Proteine

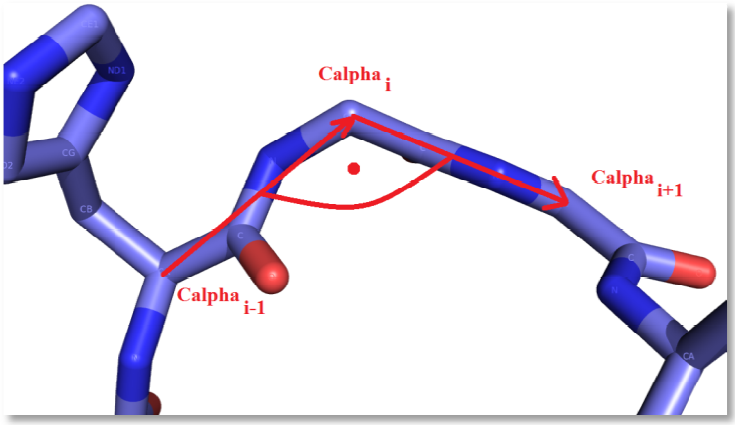
Beim Vergleich der van-der-Waals-Oberflächen beider Proteine (links: 1FS3, rechts: 1B1J), ist zu erkennen, dass die katalytisch aktiven Aminosäuren (farblich hervorgehoben) im Falle von 1B1J tiefer in der Struktur liegen, als es bei 1FS3 der Fall ist. Diese Änderungen in den Seitenkettenkonformationen könnten der Grund für die kleinen Unterschiede im Energieprofilverlauf sowie die geringere Enzymaktivität des Angiogenins sein.

5. Untersuchung von Struktur-Energie-Korrelationen

Bei der Berechnung der Energieprofile fließen verschiedenartige Informationen zusammen. Zum einen greift der Algorithmus auf die Koordinaten der einzelnen C-α- und C-β-Atome, sowie auf physikochemische Eigenschaften der betrachteten Aminosäure zu. Demnach ist das Energieprofil eine Abstraktion auf chemischer und zugleich struktureller Ebene, wodurch das Profil jedes Proteins einzigartig ist. Dieser neuartige Ansatz ermöglicht somit die Beschreibung und Klassifizierung von Proteinen auf der zweidimensionalen Ebene der Energieprofile. Um dies zu verdeutlichen, werden im folgenden Absatz Korrelationen zwischen strukturbeschreibenden Parametern und den neuen Energieprofilen abgeleitet. Im Anschluss wird eine Möglichkeit dargelegt, dieses Wissen auf unbekannte Proteine zu übertragen und deren Strukturvorhersage und Vergleich mit bekannten Strukturen zu erleichtern.

5.1 Energie-Torsionswinkel-Korrelation

Ausgangspunkt für die Untersuchung von Residuen-Energien und der Torsionswinkel-Änderung in der Proteinstruktur bildet die Hypothese, dass die Energie äquivalent zur Kraft ist, die eine Aminosäure auf die Struktur ausübt. Verläuft die backbone des Proteins in das Lösungsmittel hinein, nimmt die Wechselwirkungsenergie mit dem umgebenden Medium zu. Folglich würde die auf die backbone wirkende Kraft die Residuen zum Proteininneren drängen. Dies hätte die Ausbildung einer turn-Substruktur zur Folge. In dem Fall würde die backbone über wenige Residuen hinweg eine Winkeländerung von 180° einnehmen.
Um diesen Zusammenhang näher zu untersuchen, wurden die Proteinstrukturen auf C-α-Cα-Vektoren differenziert. Somit konnte der Torsionswinkel zwischen den C-α-Atomen berechnet werden (Abbildung 7).



ıen

Protein-

Hierfür wurde ein neuer weitaus größerer Datensatz von 4300 PDB-Dateien von ligandenfreien und einkettigen[12] Proteinen verwendet, der mit Hilfe eines in Java realisierten, vollautomatisierten WebAgent-Programmes erstellt werden konnte.
Durch Gleichung 14 lässt sich die Winkeländerung der backbone am Residuum *i* berechnen.

$$\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|} = \cos\varphi \qquad (14)$$

Dabei definiert:

- $\vec{a}$ den Vektor des C-α-Atoms von Residuum *i*-1 zum C-α-Atom der betrachteten Aminosäure *i*
- $\vec{b}$ den Vektor vom C-α-Atom der betrachteten Aminosäure *i* zum C-α-Atom des Residuums *i*+1

Die Winkeländerung wurde zur Energie des Residuums *i* aufgetragen, wodurch der in Abbildung 18 dargestellte Plot entstand.

[12] Diese Eigenschaften sind notwendig, da weitere Peptidketten oder Liganden (Ionen oder organische bzw. anorganische proteinfremde Verbindungen) das Energieprofil verfälschen.

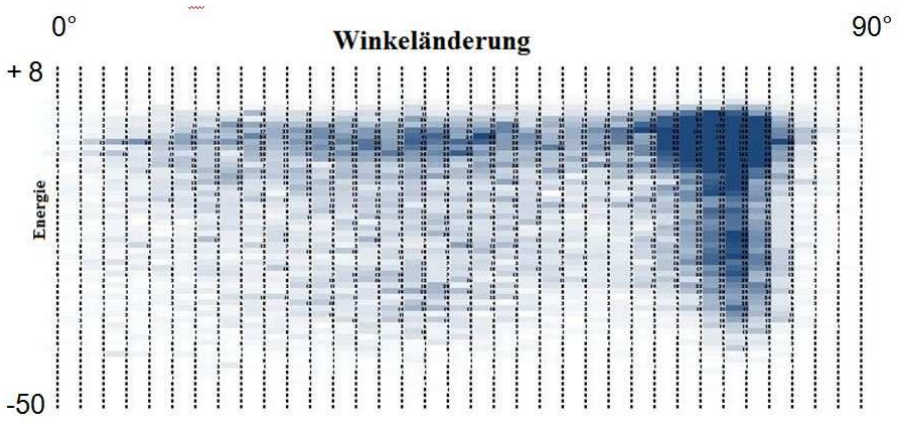


Abbildung 18: Plot-Darstellung der Energie-Torsionswinkel-Korrelation
Anhand dieses Plots lässt sich der Zusammenhang zwischen Energie und Richtungsänderungen in der Protein-Backbone ableiten. So induzieren niedrige Energien meist einen Torsionswinkel von ca. 70-75°. Residues mit höheren Energien sind dagegen indifferent.

Daraus lässt sich ableiten, dass Residuen mit einer hohen Energie in den möglichen backbone-Winkeländerungen indifferent sind. Die Größe des Torsionswinkels, den ein hochenergetisches Residue induziert, ist folglich nicht vorhersagbar. Niedrige Energien bedeuten hingegen in der Mehrzahl aller Fälle eine Richtungsänderung von 70°-75°. Für niederenergetische Residuen lässt sich demnach der Verlauf der Proteinbackbone gut beschreiben.

5.2 Energie-Sekundärstruktur-Korrelation

Der strukturelle Unterschied zwischen globulären Proteinen ist oftmals sehr groß. Das Grundprinzip des Aufbaus ist jedoch dasselbe. So liegen die Faltblattstrukturen im Inneren des Proteins, während die Helices sich schützend um diese legen. Die Coils verbinden diese Strukturen. Längere Coils sind in der Mehrzahl aller Fälle am Proteinäußeren zu finden.
Nach dem in 3.1 erläuterten Theorem ist die Energie äquivalent zur Stabilität, mit der die Aminosäure in der Struktur verankert ist. Somit ergibt sich, dass stabile, in der Struktur fest eingebettete Aminosäuren oder Sequenzabschnitte eine niedrigere Energie aufweisen. Helices und Faltblätter sind aufgrund ihrer regelmäßigen Struktur und den zusätzlichen Wasserstoffbrückenbindungen weitaus stabiler als Coils. Zudem sind Helices und Strands in ihrem Aufbau vereinheitlicht. Die φ- und ψ-Torsionswinkel

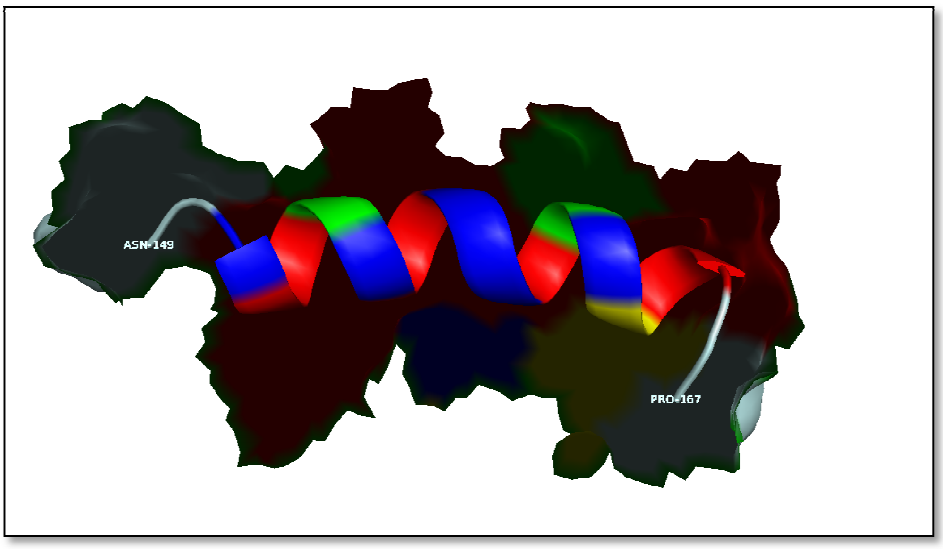
liegen in diesen Sekundärstrukturelementen in fest definierten Bereichen vor. Somit sind diese Sekundärstrukturelemente mathematisch exakt beschreibbar. Diese Manifestierung in der Struktur kommt ebenso durch den C-α_i-C-α_{i+1}-Torsionswinkel zum Ausdruck. Da nach Punkt 5.1 ein Zusammenhang zwischen Torsionswinkel und Energie vorhanden ist sollte folglich eine Korrelation zwischen der Energie eines Residuums und der Sekundärstruktur, in der sich diese Aminosäure befindet, existieren.
Für die Untersuchung dieses Zusammenhangs musste eine Quantifizierung der Energien gefunden werden. Hierfür wurden die Quantile des Energiespektrums ermittelt, wodurch eine Energieklassifizierung in vier Bereiche möglich wurde.
Nun wurden alle Energien aller Aminosäuren in Abhängigkeit ihrer Sekundärstruktur erfasst und in das jeweilige Quantil eingeordnet. Das Ergebnis dieser Auszählung zeigt Tabelle 5.

Tabelle 5: Verteilung der Aminosäuren in den Sekundärstrukturen in Abhängigkeit ihrer Energien (in 10^3)
Hier zeigen sich die unterschiedlichen energetischen Präferenzen der Aminosäuren in Abhängigkeit des Sekundärstruktur-Elementes. Die Residues in Coil-Strukturen liegen zumeist hochenergetisch vor. Aminosäuren in Faltblättern sind gehäuft niederenergetisch. In Helices lässt sich keine energetische Präferenz erkennen.

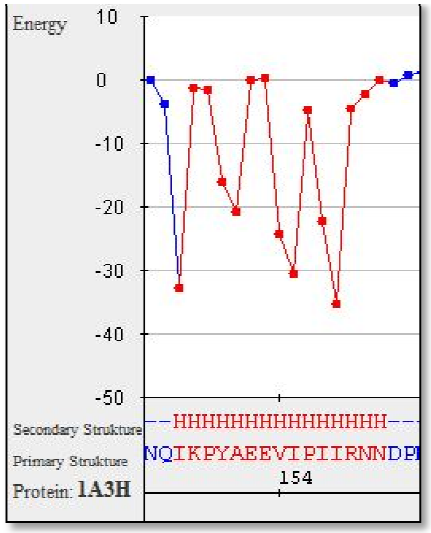
	Energie			
Sekundärstruktur-Element	**Q1** (hochenergetisch)	**Q2**	**Q3**	**Q4** (niederenergetisch)
Helix	83	80	93	97
Sheet	23	35	55	84
Random coil	128	102	83	44

Hier zeigt sich eindeutig der Zusammenhang zwischen Energie und der Sekundärstruktur einer Aminosäure. Die Energien der Sheet-Strukturen liegen in eher niederenergetischen Bereichen, während die Residuen der Coils hochenergetisch sind. Die Energien der in Helices befindlichen Aminosäuren sind dagegen indifferent. Grund hierfür könnte sein, dass diese Strukturen das Proteininnere umgeben, und somit einige Aminosäuren in das Lösungsmittel hinein gerichtet sind und folglich Lösungsmittelwechselwirkungen unterliegen. Andere Aminosäuren zeigen in das Proteininnere und sind vom umgebenden Medium isoliert. Demnach sind letztere Aminosäuren als stabil anzusehen und besitzen eine weitaus geringere Energie. Ein Beispiel für so eine Helix zeigt die Abbildung 19. Die Residues sind in Abhängigkeit

des energetischen Quantils eingefärbt[13]. Zudem ist, zur besseren Orientierung, die van-der-Waals-Oberfläche des Proteins dargestellt. Als Betrachter blickt man vom Proteininneren heraus auf die Helix. Das Energieprofil dieses Ausschnitts ist in Abbildung 20 dargestellt. An diesem Beispiel lässt sich gut erkennen, wie niederenergetische Aminosäuren in das Proteininnere zeigen, während sich die Seitenketten der hochenergetischen Aminosäuren zum Lösungsmittel orientieren.



Proteininneren
rechend ihres
nosäuren (rot
itenketten der
e.



ix

[13] Energetische Entsprechung: Rot-Q1,Gelb-Q2,Grün-Q3,Blau-Q4

5.3 Energie-SASA-Korrelation

Aus 3.2 ergibt sich die mögliche Korrelation aus der Exponiertheit einer Aminosäure zum Lösungsmittel und seiner Energie. Diese Lösungsmittelzugänglichkeit (SASA – engl.: *solvent accessible surface area*) wird durch den Quotienten aus der Lösungsmittel zugänglichen Oberfläche und der nicht zugänglichen Oberfläche beschrieben. Die SAS eines Moleküls ist durch den Abstand, ab dem zwischen einem weiteren Molekül und dem betrachteten Atom van der Waals-Wechselwirkungskräfte auftreten, definiert. Der Algorithmus zur Berechnung der SASA nutzt einen kugelförmigen 1,4-Å-Körper. Dieser repräsentiert ein Wassermolekül. Wird diese Kugel über die van-der-Waals-Oberfläche gerollt, lässt sich die SASA aller Atome einer Aminosäure, und folglich die komplette Lösungsmittelzugänglichkeit eines Residuums ermitteln.
Zur Berechnung der Lösungsmittelzugänglichkeit aller Proteine im Datensatz wurde der POPS-Server verwendet. Im Anschluss wurde die Rangkorrelation aller Energie-SASA-Wertepaare ermittelt. Der ermittelte Rangkorrelationskoeffizient betrug $r_s = 0{,}61$. Somit existiert eine starke Korrelation zwischen der Energie einer Aminosäure und dessen Lösungsmittelzugänglichkeit. Weiterhin wurden zur Visualisierung des Zusammenhangs die SASA in Zehnergruppierungen eingeteilt und die Energien auf ganzzahlige Werte gerundet. Somit konnten die Wertepaare klassifiziert, zu Gruppen zusammen gefasst und ausgezählt werden. Dieser Plot ist in Abbildung 21 zu sehen.

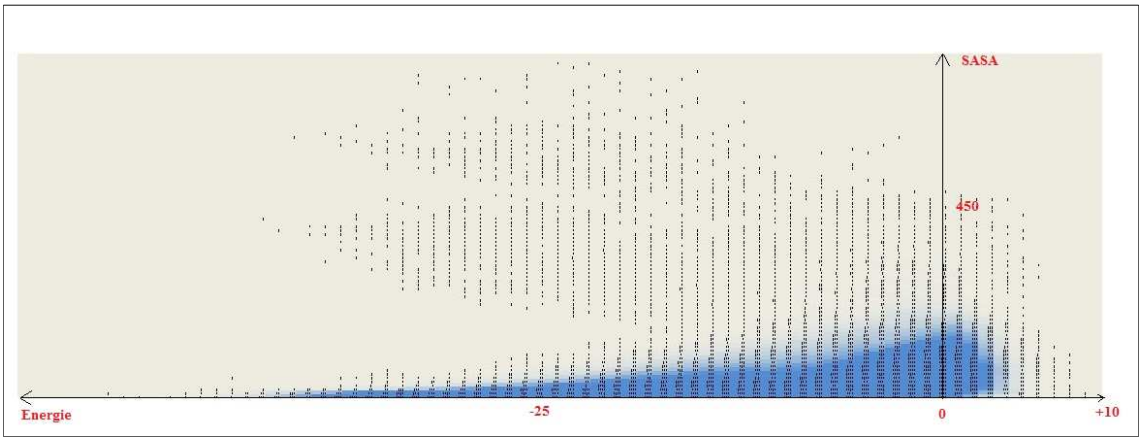


Abbildung 21: Plot-Darstellung der Energie-SASA-Korrelation

Hier zeigt sich der Zusammenhang zwischen Lösungsmittelzugänglichkeit und Energie einer Aminosäure. Niederenergetische Residues zeigen meist eine geringe SASA. Hochenergetische Aminosäuren zeigen sich indifferent.

Es ist zu erkennen, das Residues mit einer niedrigen Energie eine geringere SASA besitzen als hochenergetische Aminosäuren. Demnach ist die These der Äquivalenz von Stabilität und Energie bewiesen.
Folglich lassen sich aus dem Energieprofil konkrete Informationen zur Lösungsmittelzugänglichkeit und Stabilität ablesen.

5.4 Visualisierung der Korrelationen in $\bar{E}$-$\Delta\bar{E}$-SASA-Plots

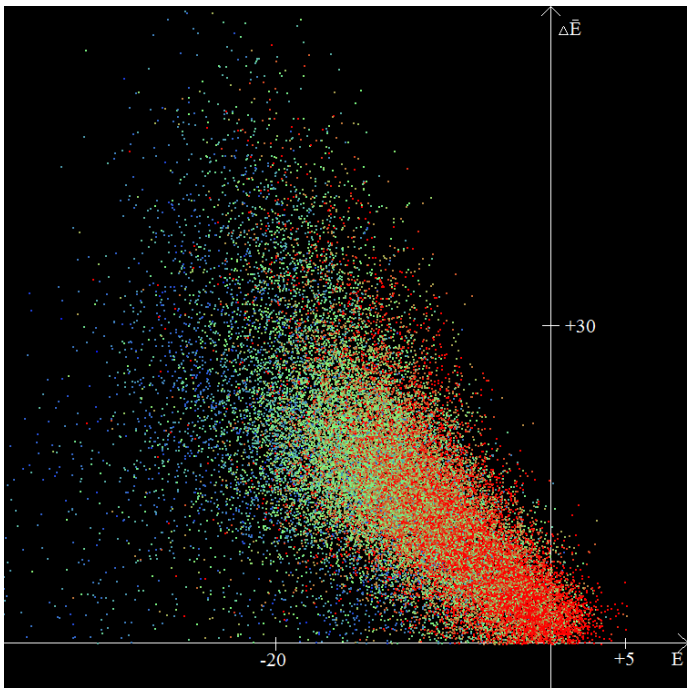
Die in 5.3 beschriebene Korrelation lässt sich mit dem in 5.2 beschriebenen Zusammenhang vereinen. Besteht eine Abhängigkeit zwischen Energie und Sekundärstruktur sowie zwischen Energie und SASA, so sollte auch zwischen der Lösungsmittelzugänglichkeit und den Sekundärstrukturelementen ein Zusammenhang feststellbar sein. Dies lässt sich folgendermaßen begründen: Wenn die Energie als Maß für Stabilität gilt und eine niederenergetische Residue in einem Strand vorliegt, so wird nach 5.2 die SASA dieser Aminosäure eher gering sein. Anders verhält es sich bei Residues in Coil-Strukturen. Diese liegen gehäuft hochenergetisch vor. Sehr oft befinden sich diese Strukturen im peripheren, für Wasser leicht zugänglichen Bereich der Proteine. Entsprechend ist die SASA dieser Residues im Schnitt höher, als bei niederenergetischen Aminosäuren.
Um diesen Zusammenhang zu visualisieren, wurde für jedes Sekundärstrukturelement ein $\bar{E}$-$\Delta\bar{E}$-SASA-Plot angefertigt. Darin repräsentiert ein Datenpunkt jeweils ein Sekundärstrukturelement. Der x-Wert definiert sich aus der durchschnittlichen Energie des Sekundärstrukturelements. Die y-Koordinate entspricht $\Delta\bar{E}$ -dem arithmetischen Mittel aller Energiebeträge in der betrachteten Sekundärstruktur. Der Energiebetrag ΔE einer Aminosäure i wird durch (15) ermittelt.

$$\Delta E_i = \left| E_i - E_{i-1} \right| \tag{15}$$

Folglich entspricht $\Delta\bar{E}$ dem Term (16).

$$\Delta \bar{E} = \frac{\sum_{i=2}^{n} \left| E_i - E_{i-1} \right|}{n-1} \tag{16}$$

Werden nun sämtliche Datenpunkte in Abhängigkeit ihrer SASA eingefärbt, erhält man einen $\bar{E}$-$\Delta\bar{E}$-SASA-Plot. Für die Färbung der Lösungsmittellöslichkeit wurde eine fließende Farbskalierung implementiert. Dabei wurde festgelegt, dass eine SASA von 0 Blau koloriert wird (RGB-Code: 0,0,255). Eine Lösungsmittelzugänglichkeit von >80 führt zu einer roten Einfärbung (RGB-Code: 255,0,0). Die Plots für jede der drei Sekundärstrukturelemente sind in den Abbildungen 22, 24 und 25 zu sehen.

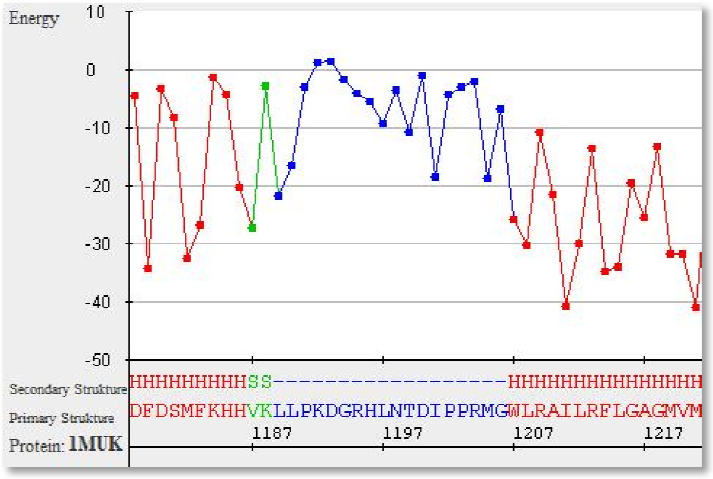


ıergieprofilverlauf und

überwiegend niedrig
Energien mit geringen
> 80). Der farbliche

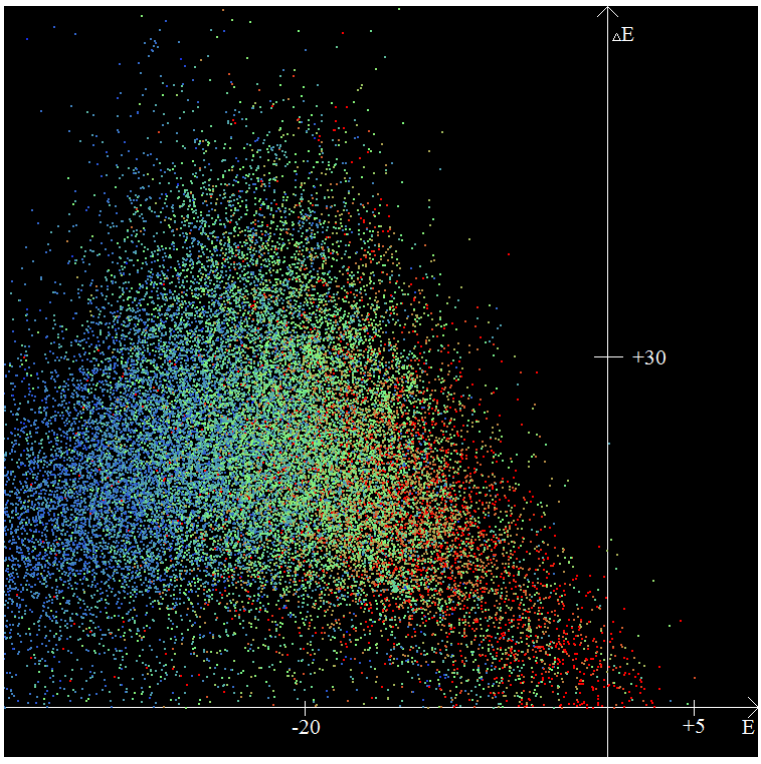
Die Abbildung 22 zeigt die Verteilung der Coil-Strukturen im $\bar{E}$-$\Delta\bar{E}$-SASA-Plot. Hier lässt sich feststellen, dass die Datenpunkte dieser Sekundärstrukturelemente sich in der Nähe des Koordinatenursprunges häufen. Demnach ist der überwiegende Teil der Coils hochenergetisch. Dies bekräftigt die in 5.2 angegebene Verteilung der Coil-Strukturen im Energiespektrum. Weiterhin sind die $\Delta\bar{E}$ überwiegend niedrig, was die Interpretation zulässt, das Coils in ihrem Energieprofilverlauf nur geringen Schwankungen unterliegen. Einen solchen typischen Profilverlauf zeigt die Abbildung 23. In diesem

Beispiel ist der Coil blau hervorgehoben. Wie ein Plateau hebt sich der Profilverlauf des Coils von den Energien der anderen Sekundärstrukturelemente ab.



lauf

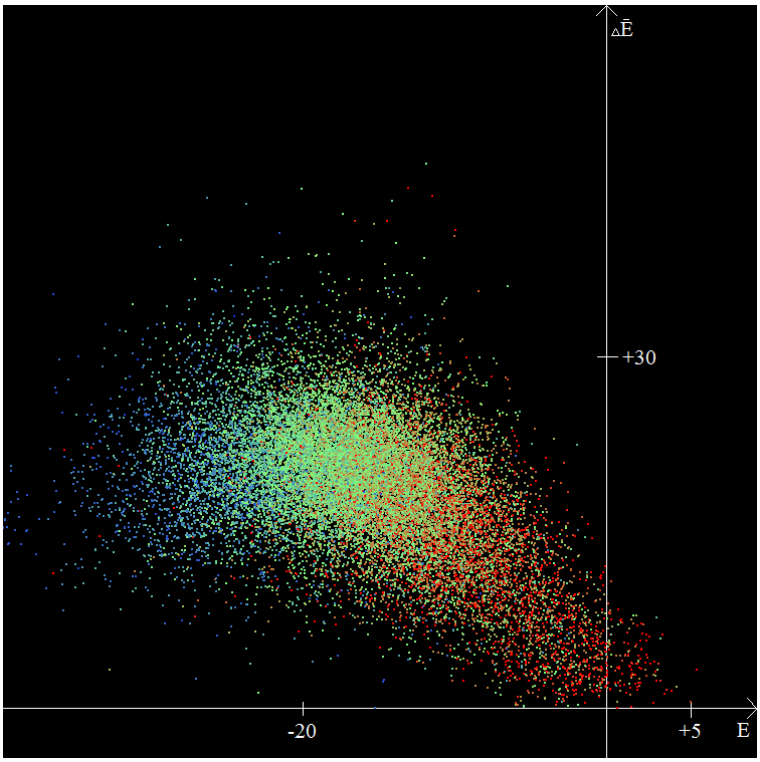
Weiterhin bestätigt der Plot in Abbildung 22 den Zusammenhang zwischen Energie und Lösungsmittelzugänglichkeit. Je höher $\bar{E}$ einer Sekundärstruktur ist, desto mehr geht die Färbung des Datenpunktes zum Rot hin über. Umso größer ist demnach die SASA.



ıstände

Dieser Zusammenhang zwischen der durchschnittlichen Energie $\bar{E}$ und der Lösungsmittelzugänglichkeit ist ebenso im $\bar{E}$-$\Delta\bar{E}$-SASA-Plot der Strand-Strukturen ersichtlich (Abbildung 24). Auch hier zeigt sich der Übergang von blauer zu roter Färbung mit zunehmender Energie und bestätigt erneut den Zusammenhang zwischen SASA und Energie.

Die in Tabelle 5 aufgezeigte energetische Verteilung der Strands ist auch im $\bar{E}$-$\Delta\bar{E}$-SASA-Plot ersichtlich. Dies kommt durch die überwiegende Verteilung der Datenpunkte in der linken Hälfte des Plots, dem niederenergetischen Bereich, zum Ausdruck. Auffällig ist zudem, dass die Schwankungen im Energieprofilverlauf, die durch $\Delta\bar{E}$ abstrahiert sind, stark variieren.



dlich,
festen
mit

Die Helices zeigen dagegen ein anderes Verhalten. Wie im $\bar{E}$-$\Delta\bar{E}$-SASA-Plot (Abbildung 25) zu erkennen ist, unterliegt der typische Energieprofilverlauf geringeren Schwankungen als es bei Strands der Fall ist. Zudem ist die in Punkt 5.2 gezeigte energetische Indifferenz ersichtlich. Ohne das eine eindeutige Tendenz erkennbar ist,

liegen die Punkte breit über das komplette Energiespektrum verteilt. Der Zusammenhang von SASA und Energie ist auch in diesem Plot anhand des fließenden Übergangs von Blau- zur Rotfärbung mit zunehmender Energie ersichtlich.

Aus diesen Plot-Darstellungen lassen sich für die drei Hauptsekundärstrukturelemente wesentliche Charakteristiken im Energieprofilverlauf heraus ableiten und in einen physikalischen Kontext bringen. Allerdings zeigt sich auch auf, dass diese Eigenschaften nicht eindeutig sind und somit die direkte Ableitung von Sekundärstrukturelementen aus einem vorhergesagten Energieprofil nur schwer möglich ist.

6. eAlign –Algorithmus zum paarweisen Vergleich zweier Energieprofile

6.1 Algorithmischer Ansatz und Implementierung

Wie im Punkt 4.2 beschrieben wurde, ist der paarweise Vergleich zweier Energien nur über eine Transformation möglich. Zudem zeigt dieser Punkt eine Lösung des Problems auf. Bei der Entwicklung von eAlign wurde dieses Prinzip erneut aufgegriffen. Nach diesem Ansatz kann nun der Energie-Score S_E für jedes beliebige Energiepaar berechnet werden. Das Verfahren der S_E-Berechnung in eAlign wurde unverändert aus dem Super-Alignment übernommen.

Um nun den Konsensus zweier Energieprofilen zu finden und aus den verwendeten Editieroperationen einen Score abzuleiten, bietet sich die Berechnung einer Optimierungsmatrix an, die durch den in Punkt 4 dargelegten Needleman-Wunsch-Algorithmus gelöst werden kann.

Folglich ist der hinter eAlign stehende Algorithmus mit dem des Super-Alignments nahezu identisch. Jedoch mussten einige Änderungen vorgenommen werden. Zum einen wurde der Algorithmus so abgeändert, dass die Bewertung von Lücken weitaus sensitiver erfolgt. Das wurde dadurch erzielt, in dem die Erweiterung einer Lücke (gap-extend $\mathcal{E}_e$) weniger Strafpunkte kostet als das Einführen einer Lücke. Dadurch wird erreicht, dass lange und kurze Lücken diskriminiert werden, wenn deren Scores signifikant unterschiedlich sind.

Diese Änderung hat auf den Algorithmus massive Auswirkungen. Zum einen wurde bei der Initialisierung der FMatrix auf die in Punkt 4.1.2 beschriebene open-gap-Methodik verzichtet. Somit ergibt sich bei der Initialisierung unter Berücksichtigung von gap-extends:

$$\begin{aligned} FMatrix(1, j) &= \varepsilon + \varepsilon_e \cdot (j-2) \\ FMatrix(i,1) &= \varepsilon + \varepsilon_e \cdot (i-2) \end{aligned} \qquad (17)$$

In der Testphase des Algorithmus wurde für $\mathcal{E}$ und $\mathcal{E}_e$ ein Score von -30 bzw. -7 festgelegt.

Der folgende Pseudo-Code zeigt, wie nun mittels des Needleman-Wunsch-Algorithmus das Alignment zweier Energieprofile und dessen Score berechnet werden kann.

```
Für i = 2, 3,… n führe aus
        Für j = 2, 3,… m führe aus

            Wenn pathMatrix[i-1,j] == „D“
                mTOP   ←  FMatrix[i-1, j] + gap-cost
            ansonsten
                mTOP   ←  FMatrix[i-1, j] + gap-extend

            Wenn pathMatrix[i,j-1] == „D“
                mLEFT   ←  FMatrix[i, j-1] + gap-cost
            ansonsten
                mLEFT   ←  FMatrix[i, j-1] + gap-extend


            mDIAG ←  FMatrix[i-1, j-1] + S_E(E_ai,E_bj)


            FMatrix[i, j] ← max(mLEFT, mDIAG, mTOP)

            Wenn FMatrix[i, j] == mDIAG
                  pathMatrix[i, j] ← “D”
            oder  FMatrix[i, j] == mLEFT
                  pathMatrix[i, j] ← “-“
            ansonsten
                  pathMatrix[i, j] ← “|”
```

Zwar lässt sich durch dieses Verfahren der Score eines Alignments berechnen und durch Backtracking der pathMatrix das Ergebnis visualisieren, jedoch ist an diesem Score (raw-Score) keine Signifikanz ablesbar.

Dieses Problem wurde durch die Berechnung eines z-Scores gelöst. Dieser Score gibt an, ob es sich bei dem erhaltenen Ergebnis um ein Zufallsprodukt handelt oder nicht. Um den raw-Score x auf Signifikanz zu testen, wird ein Energieprofil an zwei zufällig gewählten Positionen permutiert. Im Anschluss wird das oben beschriebene Verfahren erneut durchgeführt und der Alignment-Score x_P zwischen dem belassenen und permutierten Energieprofil erfasst. Für hinreichend viele n muss eine Energieprofilpermutation mit anschließender Score-Berechnung durchgeführt werden.

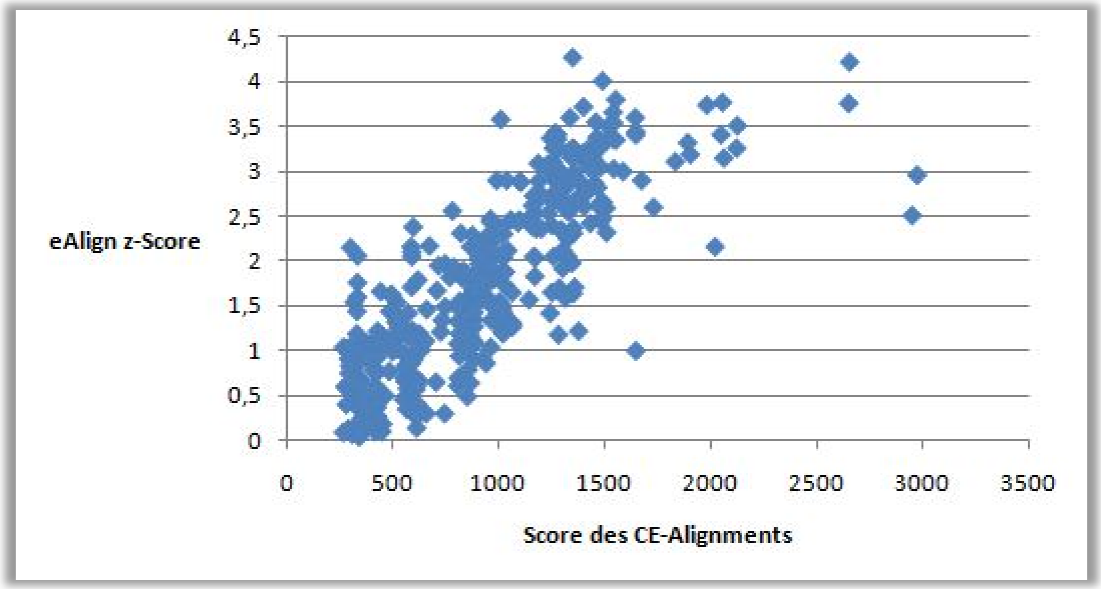
Aus den erfassten x_P lässt sich nun die Signifikanz des raw-Score ableiten. Ausschlaggebend ist hierfür der Betrag des raw-Score x vom Mittelwert aller x_P. Um diesen Betrag zu quantifizieren, muss dieser als Vielfaches von der Standardabweichung δ aller x_P verrechnet werden [7].
Demnach definiert sich der z-Score durch:

$$z - Score = \frac{X - \bar{X}_P}{\delta} \qquad (18)$$

Signifikante eAlign-Alignments besitzen einen z-Score von ca. > 1,5.

6.2 Korrelation zwischen eAlign und Strukturalignment

Für diese Untersuchung waren nur die Alignments interessant, die trotz einer geringen Sequenzidentität in ihrer Struktur und Funktion ähnlich oder identisch sind.
Hierfür wurden alle möglichen Sequenzalignments eines Datensatzes von 4300 Proteinen durchgeführt. Erfasst wurden nur die Protein-Protein-Paarungen mit einer Sequenzidentität von $20\% \leq x \leq 40\%$. Nun wurde, für jede dieser Paarungen, das Strukturalignment erzeugt und dessen Score erfasst. Dies erfolgte durch die Verwendung des CE-Algorithmus.
Für jede Sequenzidentität wurden alle Strukturalignment-Scores erfasst. Aus diesen Daten ließen sich (für jede Sequenzidentität) der dazugehörige Mittelwert und die Standardabweichung der entsprechenden Strukturalignment-Scores ermitteln und sich die Paarungen extrahieren, die einen abnorm hohen Strukturalignment-Score aufwiesen. Zu jedem dieser Wertepaare wurde im Anschluss das eAlign der Energieprofile durchgeführt und die Rangkorrelation zwischen den Scores des Strukturalignments und den z-Scores des eAligns berechnet. Das Diagramm in Abbildung 26 zeigt diesen Zusammenhang. Die Rangkorrelation beträgt $r_S = 0{,}75$. Es existiert demnach ein starker Zusammenhang zwischen dem Alignment zweier Energieprofile und dem Strukturalignment zweier Proteine.



ments

eAlign ersichtlich.

ofilvergleich. Die

6.3 Exemplarisches Ergebnis

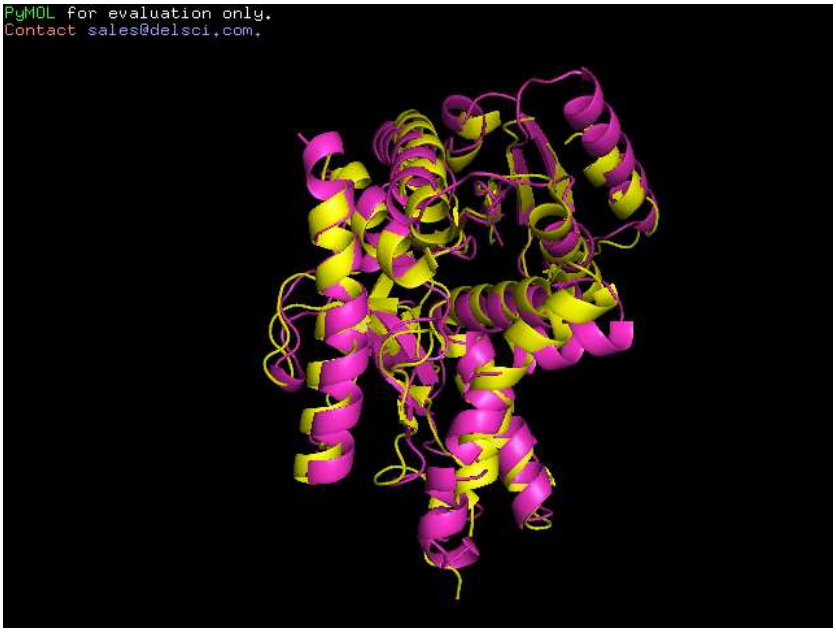
Ein besonderes Beispiel, bei dem eAlign seine Sensitivität bewiesen hat, ist das Alignment einer Malatdehydrogenase (PDB-ID: 1B8P) und einer L-Lactatdehydrogenase (PDB-ID: 1Y6J). Interessant an diesen beiden Proteinen ist die Tatsache, dass beide Moleküle sich in ihrer Funktionalität und Struktur sehr ähneln, obwohl die Sequenzidentität nur 20% und die Sequenzähnlichkeit nur 41% beträgt (siehe Abbildung 27) [33, 34].

```
SEQUENCE          1 MEMVKSRSKVAIIG-AGFVGASAAFTMALRQIANE-----LVLIDVFKEK    44
                    |.|:..:||:.| ||.:..|..|.:|......:     |.|::::..||
SEQUENCE          1 --MAKTPMRVAVTGAAGQICYSLLFRIANGDMLGKDQPVILQLLEIPNEK    48

SEQUENCE         45 A----IGEAMDINH-GLPFMGQMSLYAGDYSDVKDCDVIVVTAGANRKPG    89
                    |    .|..|:|:. ..|.:..|:.:|...:..||.||.::.....|.||
SEQUENCE         49 AQKALQGVMMEIDDCAFPLLAGMTAHADPMTAFKDADVALLVGARPRGPG    98

SEQUENCE         90 ETRLDLAKKNVMI----AKEV----TQNIMKYYNHGVILVVSNPVDIITY   131
                    ..|.||.:.|..|    .|.:    ::||.       :|||.||.:...|
SEQUENCE         99 MERKDLLEANAQIFTVQGKAIDAVASRNIK-------VLVVGNPANTNAY   141

SEQUENCE        132 MIQKWSGLPVGKVIGSGTVLDSIRFRYLLSEKLGVDVKNVHG-YIIGEHG   180
                    :..|.:.....|...:...||..|....::.|.|..|.::.. ::.|.|.
SEQUENCE        142 IAMKSAPSLPAKNFTAMLRLDHNRALSQIAAKTGKPVSSIEKLFVWGNHS   191

SEQUENCE        181 DSQLPLWSCTHIAGKNINEYIDDPKCNFIEEDKKKIAEDVKTAGATIIKN   230
                    .:....:....|.|.::.:.|:|...|     :......|...||.||..
SEQUENCE        192 PTMYADYRYAQIDGASVKDMINDDAWN-----RDTFLPTVGKRGAAIIDA   236

SEQUENCE        231 KGATYYGIA--VSINTIVETILLKNQNTIRTVGTVINGMYGI-EDVAISLP   277
                    :|.:....|  .:|:.|.:.:|.......|:|...:|.||| |.|....|
SEQUENCE        237 RGVSSAASAANAAIDHIHDWVLGTAGKWTIMGIPSDGSYGIPEGVIFGFP   286

SEQUENCE        278 SIVNSEGVQEVLQFNLTPEEEEALRFSAEQVKKVLNE-------VKNL--   318
                     :....|..:::| .|:.:     .||.|::...|||       |::|
SEQUENCE        287 -VTIENGEYKIVQ-GLSID-----AFSQERINVTLNELLEEQNGVQHLLG   329
```

1B8P bzw. 1Y6J)

he im Energieprofilverlauf

Mit einem DaliLite z-Score von 30, besitzen beide Proteine eine identische Struktur (siehe Abbildung 28).



EAlign erreichte beim Alignen der beiden Proteine einen signifikanten z-Score von 2,0. Die energetische Ähnlichkeit der beiden Proteine zeigt sich auch im ePlot. Diese spezielle Form des Dotplots wurde im Rahmen des eAlign-Programmes entwickelt und visualisiert die paarweisen Energien anhand ihres Energie-Scores. Blaue Elemente besitzen einen $S_E \geq +2$. Rote Elemente stehen für Elemente deren $S_E \geq +8$ ist. Zusätzlich ist Backtracking-Pfad grau unterlegt. Der nahezu diagonale Verlauf dieses Pfades zeigt die energetische Ähnlichkeit beider Proteine auf.

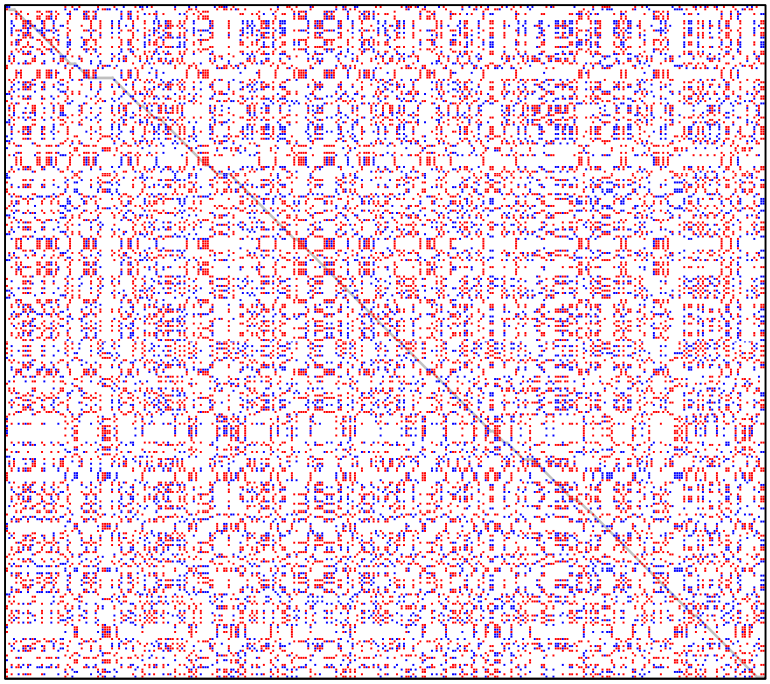


Abbildung 29: ePlot der beiden Dehydrogenasen

Diese Variation des Dotplots visualisiert alle paarweisen Energie-Scores S_E. Ein $S_E \geq +2$ ist blau hervorgehoben. Energie-Scores ≥ +8 sind rot koloriert. Zudem ist der Backtracking-Pfad grau hervor gehoben. Dieser fast diagonale Verlauf verweist auf die Ähnlichkeit beider Energieprofile.

6.4 Evaluierung des Verfahrens

Das Ziel der Evaluierung bestand darin, den Zusammenhang zwischen Strukturalignment und Energieprofilalignment zu klären. Hierfür wurde eAlign in weiteres Programm integriert. Dieses wählt aus dem Datensatz von 4300 PDB-Dateien genau 250 Proteine nach dem Zufallsprinzip. Wiederum wird aus diesen 250 Dateien ein einzelnes Protein zufällig gewählt. Mit Hilfe von eAlign sollte nun dieses eine Energieprofil in der Auswahl von 250 Proteinen gefunden werden. Zudem erzeugte das Programm für jeden Suchlauf eine Liste mit den zehn besten Treffern.

- In jedem Programmdurchlauf konnte eAlign das gesuchte Energieprofil identifizieren.

Ein Beispiel für ein erfolgreiches Ergebnis zeigte die Suche nach homologen Proteinen der C-terminalen γ-B Crystallin-Domäne (PDB-ID: 1DSL). Der Suchlauf erbrachte folgendes Ergebnis:

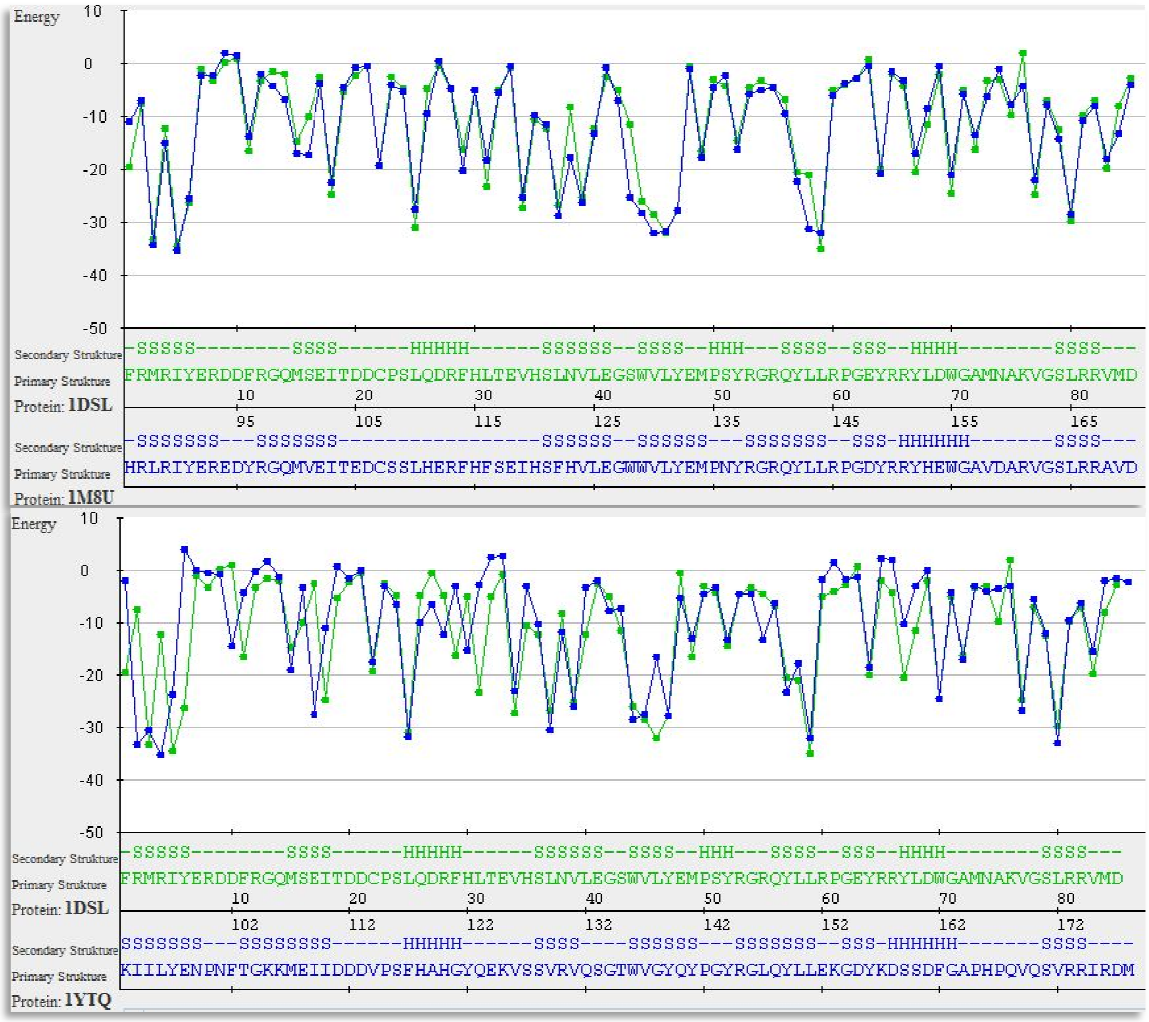
Tabelle 6: Evaluierungsergebnis im Falle des γ-B-Crystallins

Diese Tabelle zeigt das Ergebnis eines Suchdurchganges durch ein zufällig gewähltes Set von 250 Energieprofilen. Bei den zehn besten Treffern handelt es sich um Homologe des Such-templates (C-terminales Monomer des γ-B-Crystallins PDB-ID: 1DSL). Zusätzlich wurden die Ergebnisse auf strukturelle Ähnlichkeiten mit Hilfe von DaliLite untersucht sowie die biologischen Funktionen der Proteine abgeglichen.

Rang	PDB-ID	eAlign	DaliLite	Seq.-Identität	Beschreibung
1	1DSL	4,28	k.A.	100	γ-B C-terminales Monomer
2	1I5I	3,92	19,7	100	γ-B C18S mutiertes Dimer
3	1AMM	3,61	19,4	100	γ-B Dimer
4	1GCS	3,44	19,3	100	γ-B Dimer
5	1LEU	3,35	19,5	83	γ-B Dimer (theor. Modell)
6	1LFE	2,97	19,5	67	γ-C Dimer (theor. Modell)
7	1M8U	2,92	17	32	γ-E Dimer
8	1A45	2,78	16,3	75	γ-F Dimer
9	1H4A	2,74	19,2	72	γ-D R58H mutiertes Dimer
10	1LER	2,42	19,6	67	γ-C Dimer (theor. Modell)
11	1YTQ	1,87	16,6	40	β-B Dimer

Anhand dieses Beispiels lässt sich sehr gut erkennen, dass auf Grundlage des DaliLite z-Scores keine Rückschlüsse auf eventuelle Homologien gezogen werden können. Zwar weist der relativ hohe Score auf strukturelle Ähnlichkeiten hin, jedoch sind diese auf Grund der z.T. großen sequenziellen Unterschiede funktionsdivergente Proteine (siehe rechte Spalte).

Interessant ist vor allem, dass das γ-E Dimer (PDB-ID: 1M8U) trotz einer geringen Sequenzidentität den Rang 7 belegt. Ein Blick auf die Energieprofile bestätigt dies (Abbildung 30).

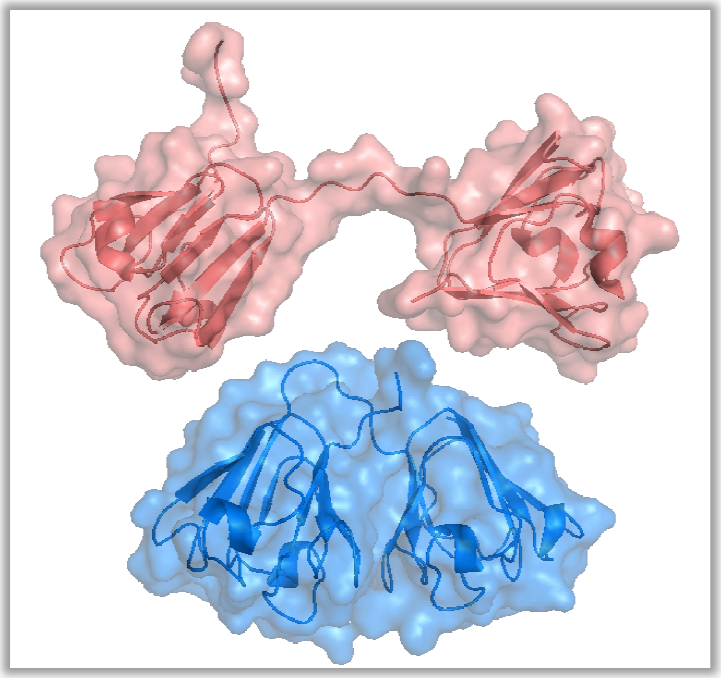


dem von 1DSL nahezu

ergetisch zeigen sich aber

Es ist eindeutig zu erkennen, dass das Energieprofil von 1M8U, bis auf wenige Abweichungen, mit dem Energieprofil von 1DSL identisch ist. Ohne Lücken einführen zu müssen, können die beiden Energieprofile zur Überlappung gebracht werden.

Anders verhält es sich beim Vergleich der Energieprofile von 1DSL und 1YTQ. Trotz einem hohen Grad an Ähnlichkeit, ist es nicht möglich die beiden Profile übereinander zulegen, ohne das Lücken eingeführt werden. Grund für diese energetischen Änderungen könnte die Änderung der räumlichen Lage der Domäne sein. Eine direkte Gegenüberstellung des β-B und γ-B Dimers zeigt den strukturellen Unterschied. (Abbildung 31).



stallin im Vergleich
einer Minderung der
-ID: 1AMM, unten)

Auffällig ist die räumliche Streckung des β-B-Dimers. Durch diese größere Distanz zwischen den Domänen, kommt es zu weit weniger Aminosäure-Kontakten als es bei γ-B-Dimeren der Fall ist. So zeigt der CMA-Plot von 1YTQ (zusehen in Abbildung 32, links) eindeutig die geringere Anzahl an van-der-Waals-Kontakten [36]. Zum direkten Vergleich dient der CMA-Plot eines γ-B-Dimers auf der rechten Seite der Abbildung 32.

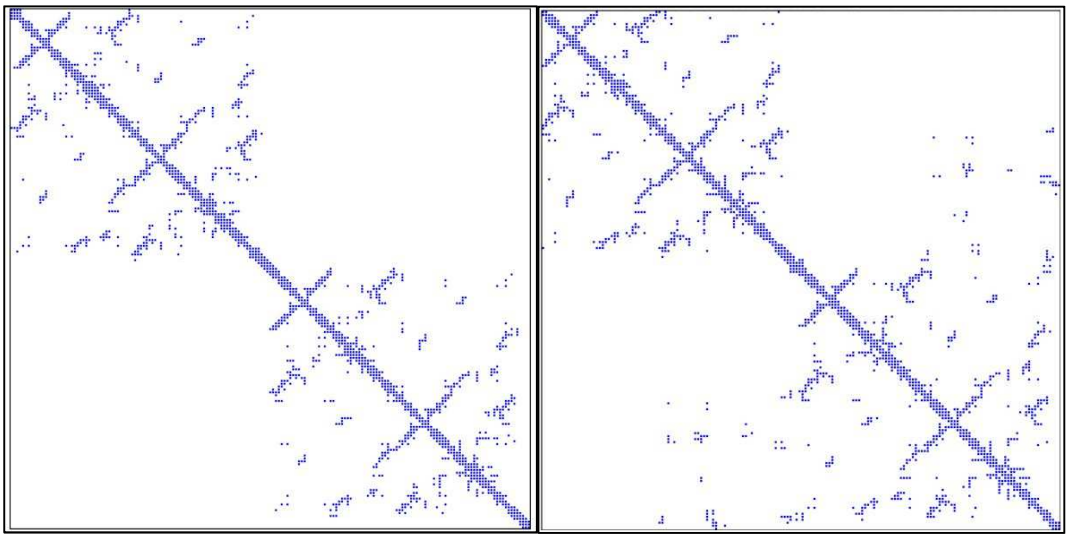


Abbildung 32: CMA-Plots zweier Crystalline [35]

Diese Plots visualisieren intermolekulare van-der-Waals-Kontakte. Das strukturell gestreckte β-B-Dimer (links) zeigt weniger Wechselwirkungen als das unveränderte γ-B-Crystallin. Dies bewirkt entsprechende Änderungen im Energieprofilverlauf (siehe Abbildung 31).

Diese geringere Anzahl an Kontakten führt dazu, dass es, nach Punkt 3.2, zu einer Änderung des Energieprofilverlaufes kommt, da entsprechende Wechselwirkungen zwischen den Monomeren verloren gehen.

6.5 Interpretation der Ergebnisse

Aus diesen Ergebnissen lassen sich einige Aspekte ableiten. Zum einen konnte mit Hilfe von eAlign der Zusammenhang zwischen Struktur und Energieprofil bewiesen werden. Jedoch zeigt sich anhand der Rangkorrelation von $r_S = 0{,}75$, dass zwischen strukturell identischen Proteinen mit einer geringen Sequenzidentität nicht unbedingt eine Ähnlichkeit zwischen den Energieprofilen bestehen muss. Begründen lässt sich das durch den in den Punkten 3.2 und 3.3 aufgeführten Zusammenhang von Energieberechnung und chemischer Zusammensetzung des Proteins. Verändert sich die Aminosäuresequenz in der 8Å-Umgebung einer Residue, so ändert sich auch dessen Energie. Fließen somit Sequenz- und Strukturinformationen zu einem Energieprofil zusammen, kann das statische Modell zu einer Verzerrung des Energieprofils führen. Strukturell ähnliche Proteine können sich in ihrem Energieprofil unterscheiden.

Somit liegt die Vermutung nahe, dass ein Energieprofil-Alignment einen Vergleich der Proteinfunktionalität darstellt. Das reine Strukturalignment, wie es DaliLite oder CE ausführen, kann es nicht ersetzen, es aber um eine viel versprechende Komponente erweitern.

7. Vorhersage von Energieprofilen unter Verwendung des GOR-Algorithmus

Das folgende Kapitel beschäftigt sich mit der Möglichkeit, auf Basis einer Aminosäuresequenz, ein Energieprofil vorherzusagen. Ziel ist es, aus dem vorhergesagten Energieprofil Rauminformationen abzuleiten, sowie mögliche strukturell ähnliche Proteine, trotz geringer Sequenzidentitäten, anhand bekannter Energieprofile zu identifizieren. Die Grundlage für dieses Vorgehen liefert der GOR-Algorithmus.

7.1 Klassischer GOR-Algorithmus

Ein in der Literatur gut beschriebener Algorithmus zur Sekundärstrukturvorherzusage von Proteinen, ist die GOR-Methode (Garnier-Osguthorpe-Robson). Er berücksichtigt die sequenzielle Umgebung einer Residue *i*. Der Algorithmus unterteilt sich in zwei getrennte Schritte. Zum einen muss, mit Hilfe eines ausreichend großen Datensatzes, eine GOR-Vorhersage-Statistik geschaffen werden. Der zweite Schritt -die durch den GOR-Algorithmus beschriebene Vorhersagemethodik -zieht diese erstellte Statistik zu Vorhersage der Sekundärstrukturelemente heran. Das Grundprinzip des Algorithmus lässt sich folgendermaßen zusammen fassen:

- Die Sekundärstruktur der Aminosäure *i* ist abhängig von der sequenziellen Umgebung
- Diese Umgebung wird durch die Aminosäuren $i \pm 8$ beschrieben
- Beim Erstellen der Statistik werden die sequenziellen Konstellationen in der Umgebung in Verbindung mit der Sekundärstruktur der Aminosäure *i* erfasst
- Bei der Prognostizierung der Sekundärstruktur an der Stelle *i,* wird die sequenzielle Umgebung um die Aminosäure *i* analysiert und mit der Statistik abgeglichen
- Aus diesem Abgleich kann auf die Sekundärstruktur der Aminosäure *i* geschlossen werden

Der klassische GOR-Algorithmus besitzt eine Vorhersagegenauigkeit von ca. 60-65% [14].

7.2 Abwandlung des Algorithmus

Um die Vorhersage von Energieprofilen zu ermöglichen, mussten die im klassischen GOR-Algorithmus beschriebenen Parameter abgewandelt werden. Zum einen wurde die Länge der betrachteten Umgebung geändert. Der Grund hierfür liegt darin, dass die Energie einer Residue sich aus den Wechselwirkungsenergien mit den Aminosäuren mit einem Abstand $r \leq 8$ Å (siehe Punkt 3.2) vom betrachteten Residuum zusammensetzt. Die benutzte Fensterlänge wurde aus der, unter Punkt 3.3 dargelegten Untersuchung abgeleitet. Nach dieser Untersuchung nach unterliegt eine Residue *i* Wechselwirkungen auf *globaler* und *lokaler* Ebene. So wird im Punkt 3.3. beschrieben, das ca. $i \pm 3$ sequenziell benachbarte Aminosäuren die Residue *i lokal* beeinflussen. Folglich muss, zum relativ exakten Erfassen des *lokalen* Einflusses, die zu betrachtende Umgebung des GOR-Algorithmus $i \pm 3$ Aminosäuren betragen. Der *globale* Einfluss, wird durch die räumliche Struktur des Proteins bestimmt und kann anhand dieses Modells weder vorhergesagt noch einbezogen werden.
Weiterhin wurde die in 5.2 beschriebene Quantifizierung der Energie angewendet. Die Einteilung in vier energetische Intervalle hat sich beim Testen der Vorhersagegenauigkeit als günstigste Klassifizierung erwiesen, da dadurch das bestmöglichste Verhältnis von Genauigkeit und Informationsgehalt vorliegt.

7.3 Vorhersage von Energieprofilen

Der klassische GOR-Algorithmus erzeugt eine Vielzahl von Wahrscheinlichkeitsmatrizen, die bei der anschließenden Vorhersage genutzt werden. Der Umgang mit diesen Matrizen ist zeit- und rechenintensiv. Diese Problematik wurde durch den Einsatz von HashMaps umgangen.
Diese Datenstruktur ist eine spezielle Form der Liste, wobei die Besonderheit darin liegt, dass der Zugriff auf die Elemente in dieser Struktur nur durch Angabe eines Keys erfolgt. Dieser Schlüssel wird für die Adressierung des Zeigers in dieser Liste benötigt

(*hashing*), wodurch der Zugriff auf ein Element in dieser Liste in einer Zeitkomplexität von O(1) erfolgt.
Der Algorithmus ist so gelöst, das die betrachtete sequenzielle Konstellation den HashKey bildet und so zeitnah die Elemente der Liste bearbeitet werden können.

Das folgende Flowchart verdeutlicht den algorithmischen Ablauf zum Erstellen der Vorhersagestatistik.

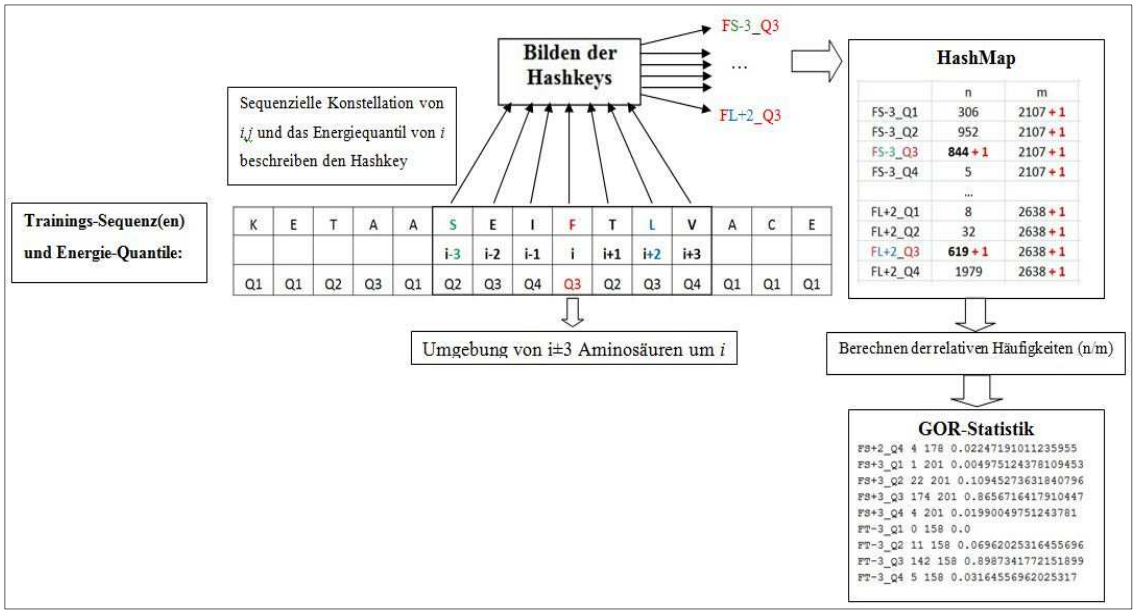


Abbildung 33: Erstellen der GOR-Statistik mit Hilfe von HashKeys

- es werden alle Konstellationen *i,j* in der Umgebung erfasst
- *n* entspricht der Anzahl aller Konstellationen von *i,j* bei denen sich *i* im energetische Zustand *k* befindet
- *m* entspricht der Gesammtanzahl der Konstellation *i,j*
- die erfasste relative Häufigkeit n/m ist Äquivalent zur Wahrscheinlichkeit, dass, in der Konstellation *i,j*, die Aminosäure *i* das Energiequantil *k* einnimmt

Mit Hilfe dieser Wahrscheinlichkeiten, lässt sich das Energiequantil von *i* und somit das komplette Energieprofil einer Input-Sequenz vorhersagen. Das folgende Flowchart demonstriert exemplarisch diesen Vorgang.

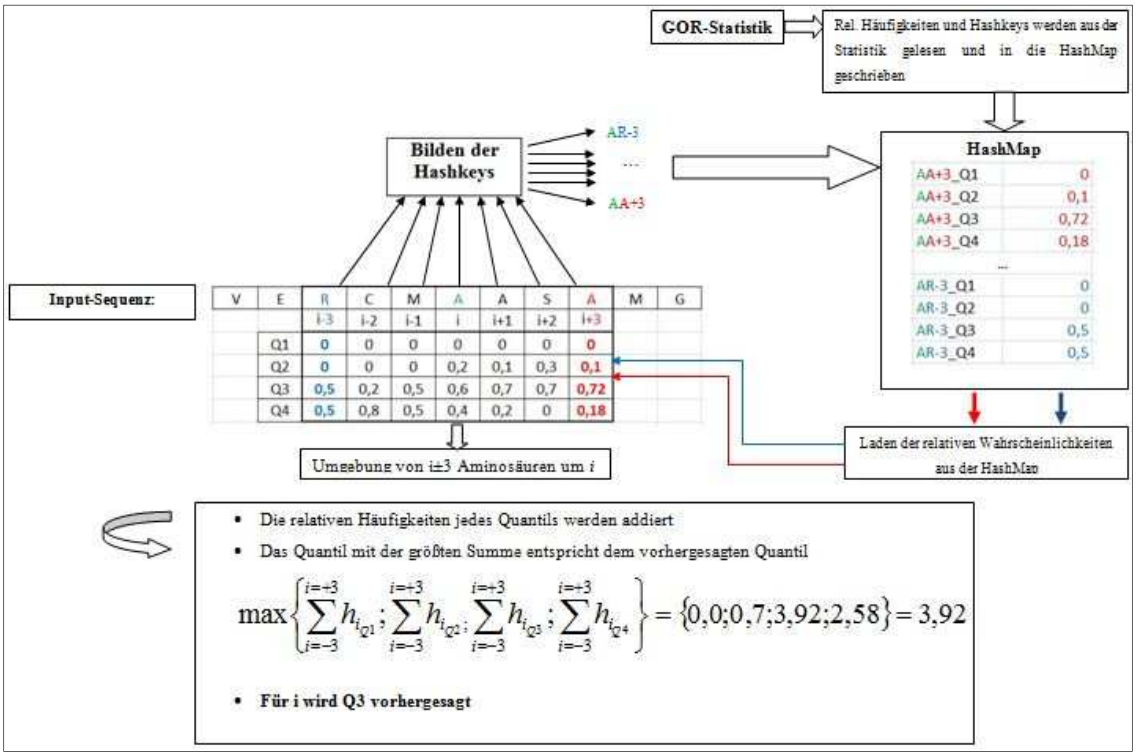


Abbildung 34: Darstellung der Energie-Quantil-Vorhersage

- die GOR-Statistik wird in eine HashMap überführt
- die sequenzielle Konstellation *i,j* bildet den HashKey
- mit diesem Key kann auf die in der HashMap eingetragenen Wahrscheinlichkeiten zugegriffen und der wahrscheinlichste energetische Zustand der Aminosäure *i* ermittelt werden

Da der Algorithmus für eine Quantil-Vorhersage für eine Residue *i* die sequenzielle Umgebung von $i \pm 3$ Aminosäuren benötigt, kann für die sechs terminalen Aminosäuren (drei Aminosäuren pro Terminus) keine Prognostizierung getroffen werden. Für eine Input-Sequenz mit einer Länge von 60 Aminosäuren bedeutet dies, dass 10% der Informationen bei der Vorhersage verloren gehen.

7.4 Vorhersagegenauigkeit

Um die Vorhersagegenauigkeit abzuschätzen, wurde eine Prognostizierung für alle Proteine des Datensatzes durchgeführt und mit den tatsächlichen Energieprofilen abgeglichen. Eine solche Vorhersage zeigt die Abbildung 16.

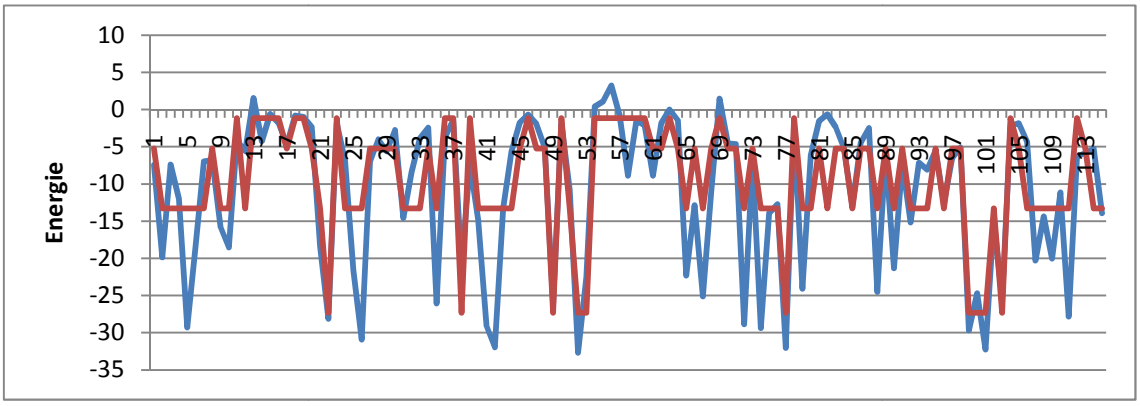


Abbildung 35: Vergleich eines vorhergesagten Energieprofils (rot) mit den realen Werten (blau)

Auf den ganzen Datensatz bezogen, ergab sich für die Vorhersagegenauigkeit folgende Statistik:

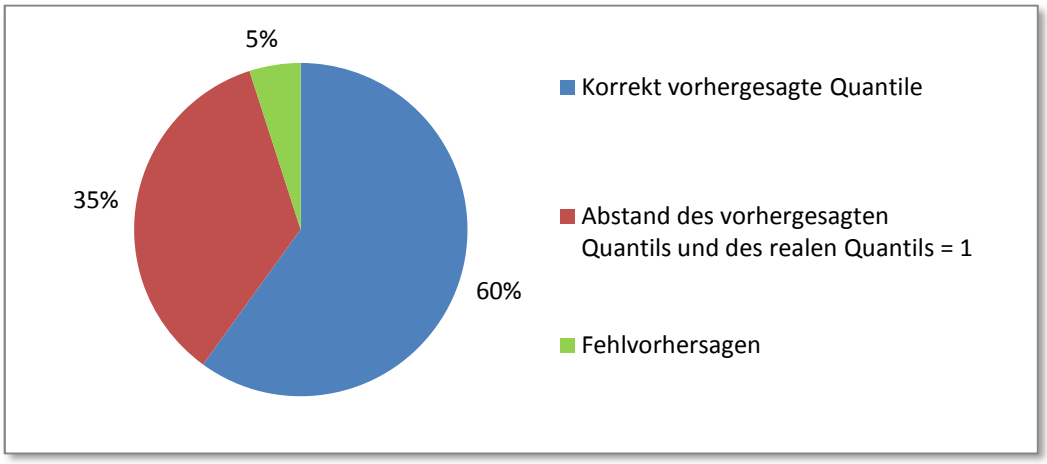
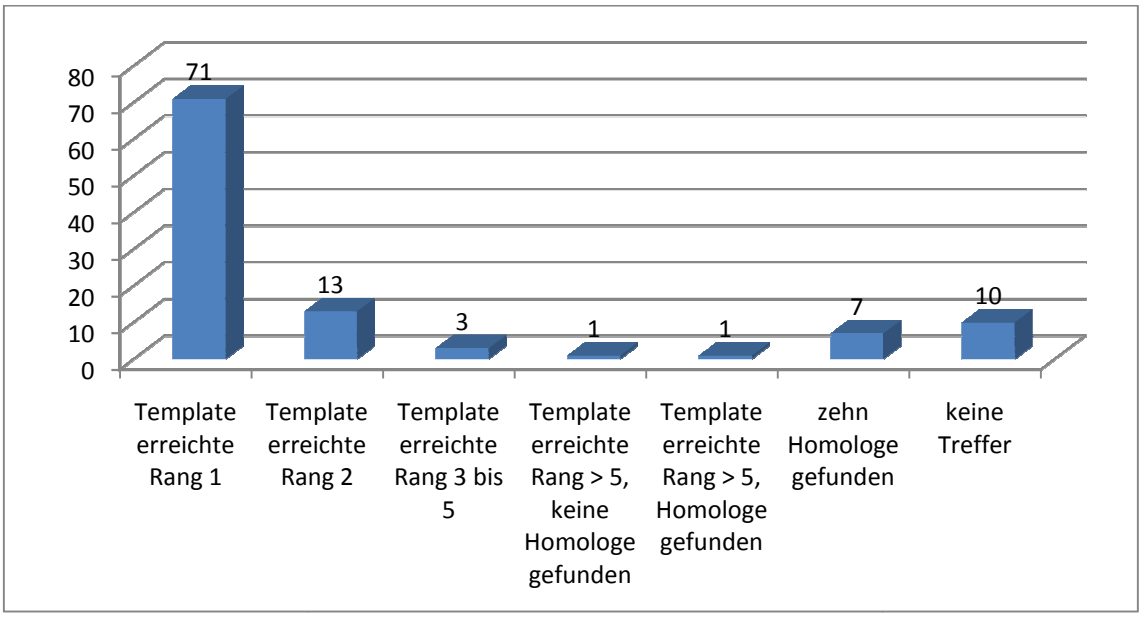


Abbildung 36: Die Vorhersagegenauigkeit des Verfahrens im Überblick

Diese Genauigkeit garantiert eine korrekte Prognostizierung des Profilverlaufs von durchschnittlich 95%. Zudem ist die Aussage zulässig, dass die lokalen Einflüsse den Energiewert zu 60% bestimmen. Die übrigen 40% werden, aufgrund der Proteinfaltung, durch sequenzferne Aminosäuren induziert.

7.5 Cross Validation von eGOR

Für die Evaluierung des Verfahrens, wurde eine leicht veränderte Methodik angewandt, wie sie in Punkt 6.4 beschrieben wurde. So wurde zu Beginn des Suchlaufes ein Protein zufällig ausgewählt und im Anschluss, auf der Basis seiner Sequenz, mit Hilfe von eGOR das Energieprofil vorhergesagt. Im Anschluss wurde der komplette Datensatz nach diesem Template-Profil oder nach ähnlichen Energieprofilen durchsucht. Hierfür wurde eAlign verwendet. Zudem wurde für jeden Suchdurchgang eine Liste mit den zehn besten Treffern erstellt. Insgesamt durchlief das Programm 106 Durchgänge, wobei jede Top-10-Liste erfasst wurde. Im Anschluss wurden alle Einträge in allen Listen mit Hilfe CE-Algorithmus auf strukturelle Identitäten mit dem Template-Protein untersucht. Zudem wurden die Sequenzidentitäten erfasst. Dadurch konnte geklärt werden, ob neben den Templates auch homologe Proteine identifiziert wurden.
Das Resultat zeigt das Diagramm in Abbildung 37.



vählt und mittels eGOR
gang durch alle 4300
Insgesamt wurden 108

Es ist also ersichtlich, dass in rund 66% aller Fälle das Template-Profil identifiziert werden konnte, indem es den höchsten Score lieferte. In sieben Durchläufen erreichte

das Template nicht die Liste der besten zehn Treffer, jedoch konnten dabei zehn homologe Proteine detektiert werden. Der Grund dafür, dass in diesen zehn Fällen eAlign das Template-Profil nicht finden konnte, liegt darin, dass eGOR Energieprofile mit nur vier Werten erzeugt. Folglich können Energieprofile, die dem Template ähnlich sehen, besser an das vorhergesagte Energieprofil fitten. Der eAlign-Score dieser Alignments liegt logischerweise höher.

In den Fällen, in denen weder das Template oder homologe Proteine gefunden werden konnten, beträgt die durchschnittliche Länge der Inputsequenzen rund 69 Aminosäuren. Grund hierfür könnte sein, dass, gemäß Punkt 7.3, der Verlust von Informationen als Folge der Energieprofilvorhersage die Identifizierung des Templates erschwert.

8. Anwendungsmöglichkeiten und Ausblick

Der in dieser Arbeit gezeigte Informationsgehalt von Proteinenergieprofilen ermöglicht eine neuartige Methodik in der Proteinanalytik. So könnten die Energieprofile von Proteinbeständen (Proteine mit bekannter Struktur) berechnet und in Datenbanken überführt werden. Mit Hilfe von heuristischen Algorithmen wäre es möglich, Proteine mit ähnlichen Energieprofilen zu identifizieren. Da diese Profile ein hohes Maß an Struktur- und physikochemischen Informationen beinhalten, aber als zweidimensionaler Vektor vorliegen, können diese in einer ähnlichen Zeitkomplexität verglichen werden, wie es bei Primärstruktur-Alignments der Fall ist. Liegt von einem Protein ausschließlich dessen Sequenz vor, kann mit Hilfe der GOR-Vorhersage das Energieprofil prognostiziert, ähnliche Energieprofile von aufgeklärten Proteinen gesucht, und so Rückschlüsse bezüglich der potentiellen Struktur, Funktionalität und Enzymaktivität getroffen werden.

Zudem könnten strukturgebende Motive eindeutiger identifiziert werden. Es besteht die Möglichkeit die aktiven Zentren der Proteine auf energetischer Basis zu beschreiben, wodurch Rückschlüsse bezüglich deren Aktivität getroffen werden könnten.

Weiterhin ermöglicht das Verfahren, die globalen Einflüsse von Punktmutationen näher zu beschreiben. Ebenso könnte neues Wissen auf diesem Gebiet dazu beitragen, die Leistungsfähigkeit und Präzision von Homology-Modelling-Servern zu optimieren. Mit weiteren Erkenntnissen wäre man in der Lage, die Lücke zwischen Sequenz und Struktur weiter zu schließen, wodurch eines Tages die vollständige Beschreibung von Proteinfaltung und Funktion ermöglicht wird.

Literaturverzeichnis

[1] URL:<http://www.iheinrich.com/blog/?p=1191>

[2] Selzer, Marhöfer, Rohwer.: *Angewandte Bioinformatik – Eine Einführung*, 1. Aufl. Berlin Heidelberg: Springer 2004

[3] Perryman et al, *Fragment-Based Screen Against HIV Protease*, Chem Biol Drug Des 2010; 75: 257–268

[4] F.Dressel, A. Marsico, A. Tuukkanen, M. Schroeder and Labudde, D.; *Understanding of SMFS barriers by means of energy profiles*; Proc. GCB

[5] Müller-Esterl, W.: *Biochemie*, 1.Aufl. Berling: Spektrum,2004

[6] Pollard, T. D.; Earnshaw, W. C: *Cell Biology*, 2. Aufl. Berlin: Springer,2007

[7] Lesk, A. M.: *Bioinformatik-Eine Einführung*, 1. Aufl.Berlin: Spektrum, 2002

[8] Merkl,R.;Waack,S.: *Bioinformatik Interaktiv*, 1. Aufl. Weinheim:WILEY-VCH, 2003

[9] Needleman, S; Wunsch,C.: *A Gerneral Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins*; J. Mol. Bwl. 1970; 48: 443-453

[10] PDB-ID *1FS3*, URL:<http://www.pdb.org/pdb/explore/explore.do?structureId=1FS3>

[11] PDB-ID *1B1J*, URL:<http://www.pdb.org/pdb/explore/explore.do?structureId=1B1J>

[12] Raines, R. T.: *Ribonuclease A*, Chem. Rev. 1998; *98:* 1045−1065

[13] PDB-Sum-Entry *1DSL*, URL:< http://www.ebi.ac.uk/thornton-srv/databases/cgi-in/pdbsum/ >

[14] Garnier, J.; Osguthorpe, D.J.; Robson, B.: *Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins*; Journal of Molecular Biology (1978); 120: 97-120

[15] Hertrich, M.: *Membran-Aufbau*, URL:<http://www.cellmicrocosmos.org/download/cm2_1/Membran-Aufbau.pdf>

[16] Protein Data Bank, URL:<http://www.pdb.org/pdb/search/advSearch.do>

[17] URL:<info@pdbwiki.org, http://pdbwiki.org/index.php/Main_Page>

[18] URL:<www.scheffel.og.bw.schule.de/faecher/science/biologie/proteine_enzyme/1protein/ami no.gif>

[19] URL:<http://www.dsimb.inserm.fr/~debrevern/VENN_DIAGRAM/aa_venn_diagram.png>

[20] URL:<http://www.zum.de/Faecher/Ch/Saar/Handrchg/Kl11/h11_55_1.gif>

[21] modifiziert nach Lesk, A. M.: *Bioinformatik-Eine Einführung*, 1. Aufl.Berlin: Spektrum, 2002

[22] Vorlesungsreihe der Universität Kiel, URL:<http://www.unikiel.de/Biochemie/unterverzeichnisse/teaching/undergrdent/pdfachim/Proteine3.PDF>

[23] *Cryus Leventhal*, URL:<http://www.columbia.edu/cu/biology/faculty/chasin/cyrus.html>

[24] Wagner, R.: *Alzheimer BSE und freie Enthalpie: Proteinfaltung und neurodegernerative Erkrankungen*; URL:<http://www.google.de/url?sa=t&source=web&ct=res&cd=1&ved=0CAcQFjAA&url=http%3A%2F%2Fwww.physik.uni-osnabrueck.de%2Fdoc%2FProteinfaltungundKrankheiten.ppt&rct=j&q=uni+osnabr%C3%BCck+proteinfaltung&ei=PXJcS6SyOoOXsQaVzZzLAw&usg=AFQjCNF_NpR906Sb-AVdXRqtU3AnT2enPg>

[25] URL:<http://wiki.cmbi.ru.nl/images/5/5d/Phipsi.jpg>

[26] URL:<http://www.chemgapedia.de/vsengine/media/vsc/de/ch/8/bc/faltung/gif_jpgs_pdb/ golf4_swf_altref_001.jpg>

[27] Smith, J. C.: *Streifzug durch die Energielandschaft der Proteine*, URL:<http://www.uni-heidelberg.de/presse/ruca/ruca04-01/4.html>

[28] Dressel, F.: Sequenz, *Energie,Struktur-Untersuchungen zur Beziehung zwischen Primär- und Tertiärstruktur in globulären und Membran-Proteinen*, Dissertationsschrift

[29] URL:<http://www.chemgapedia.de/vsengine/vlu/vsc/de/ch/8/bc/vlu/modelling/sekundaerstrukturvorhersagen.vlu/Page/vsc/de/ch/8/bc/modelling/chou_fasman.vscml.html>

[30] NCBI, URL:<http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt>

[31] Likic´,V.: *The Needleman-Wunsch algorithm for sequence alignment*; URL:<http://www.ludwig.edu.au/course/lectures2005/likic.pdf>

[32] *Wasser und Energie*: URL:<http://www.wissenschaft-technik-ethik.de/ wasser_energie.html#kap02>

[33] PDB-ID: 1B8P URL:<http://www.pdb.org/pdb/explore/explore.do?structureId=1B8P>

[34] PDB-ID: 1Y6J URL:<http://www.pdb.org/pdb/explore/explore.do?structureId=1Y6J>

[35] CMA: Contact Map Analysis URL:<http://ligin.weizmann.ac.il/cma/>

[36] Sobolev, V; et al.: Automated analysis of interatomic contacts in proteins; Bioinformatics (1999); 15: 327-332

Danksagung

Diese Arbeit wäre ohne eine Vielzahl von Personen, die mich unterstützten, nicht möglich gewesen. Ich möchte meinem Prof. Labudde für die Unterstützung in wissenschaftlichen Fragen und für die Betreuung während dieser Arbeit danken. Die meisten Methoden und Ergebnisse, die in dieser Arbeit dargestellt sind, hätten ohne seine Anregungen und Ideen niemals einen Weg auf unbeschriebenes Papier gefunden. Zudem möchte ich meinem Zweitbetreuer Dipl.-Inf. Daniel Stockmann für seine Hilfestellungen und Lösungsansätze einen großen Dank aussprechen. Allzu oft hat sein geschulter Blick auf fehlerhaften und kryptisch geschriebenen Quellcode zum schnellen Auffinden von Fehlern geführt. Mein Dank gilt auch in diesem Zusammenhang Stefan Schildbach für die zahlreichen fruchtenden informatischen und programmiertechnischen Diskussionen. Weiterhin möchte ich mich an dieser Stelle bei Riccardo Brumm für seine mühevolle Implementierung einer Visualisierungs-Software für Energieprofile bedanken. Diese hat zum Verständnis des Problems bei getragen. Zudem hat dieses Software-Tool über verschiedenste Abbildung seinen Weg in diese Arbeit gefunden.
Meine höchste Wertschätzung gebührt auch dem Rest der Projektgruppe. Zahlreiche inspirierende Diskussionen gaben mir immer wieder zusätzliche Anregungen. Ein großer Dank gilt auch meinem Vater, der sich die Zeit für eine gründliche Suche nach Druckfehlern nahm. An dieser Stelle gebührt mein Dank dem Rest meiner Familie. In schwierigen Situationen haben sie mir immer Rückhalt gegeben.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur angefertigt habe.

Claußnitz, den 22.08.2010 Florian Heinke